2006

Distance-based protein structure modeling

Di Wu
Iowa State University

Distance-based protein structure modeling

by

Di Wu

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Bioinformatics and Computational Biology; Applied Mathematics

Program of Study Committee:
Zhijun Wu, Co-major Professor
Robert Jernigan, Co-major Professor
Drena Dobbs
Kai-Ming Ho
Vasant Honavar

Iowa State University

Ames, Iowa

2006

Graduate College

Iowa State University

This is to certify that the doctoral dissertation of

Di Wu

has met the thesis requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Co-major Program

Signature was redacted for privacy.

For the Co-major Program

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Protein structure modeling can be studied based on the knowledge of interactions or distances between pairs of atoms, which is so-called distance-based protein structure modeling and this field includes problems of structure determination and refinement as well as analysis of protein dynamics. The distances for certain pairs of atoms in a protein can often be obtained based on our knowledge on various types of bond-lengths and bond-angles or from physical experiments such as nuclear magnetic resonance (NMR). The coordinates of the atoms and hence the protein structure can then be determined by using the known distances. However, it requires the solution of a mathematical problem called the distance geometry problem, which has been proven to be computationally intractable in general. On the other hand, due to insufficient distance data such as nuclear overhauser effect (NOE) data in NMR, the protein structures determined by conventional techniques usually are not as accurate as desired. Therefore, the uses of such protein structures in important applications including homology modeling and rational drug design have been severely limited. In this work, we have developed several efficient algorithms including theories for the solution of the distance geometry problem using a geometric build-up algorithm. We also introduced a knowledge-based method for protein structure refinement, in which we constructed a dedicated structural database for protein inter-atomic distance distributions and derived so-called mean force potentials to refine NMR-determined protein structures. We have participated in CASPR competition regarding comparative models and reported some substantial improvement using mean force potentials. Finally, an efficient and simple method called Local-DME calculations has been developed to study protein dynamics of NMR ensembles specifically.

CHAPTER 1. GENERAL INTRODUCTION

Introduction

Proteins are essential to all kinds of life. Usually a biological system has a great number of proteins, each with a specific role in the system. A protein is a polypeptide chain, which contains hundreds of amino acids and thousands of atoms. In nature, there are about 20 different types of amino acids. A protein sequence and properties of amino acids determines its tertiary structure as well as its function. Therefore, knowledge of structures is very crucial for understanding and study of protein dynamics and functions.

In general, protein structure modeling can be studied based on the knowledge of interactions or distances between pairs of atoms, which is so-called distance-based protein structure modeling and this field includes problems of structure determination and refinement as well as analysis of protein dynamics.

Often the distances between certain pairs of atoms in a protein can be obtained based on our knowledge of various types of bond-lengths and bond-angles or from physical experiments such as nuclear magnetic resonance (NMR). The coordinates of the atoms and hence the protein structure can then be determined by using the known distances. However, this approach requires the solution of a mathematical problem called the distance geometry problem, which has been proven to be computationally intractable in general [4]. Therefore, for a large system, developing an efficient algorithm which is numerically stable becomes urgent and necessary. Due to insufficient distance constraints obtained from experiments, such as nuclear overhauser effect (NOE) data in NMR, the protein structures determined by conventional techniques usually are not as accurate as desired. The uses of such protein structures in important applications including homology modeling and rational drug design hence have been severely limited. Developing an efficient and reliable refinement technique is necessary, and this need becomes urgent with more and more structures determined, as the CASP prediction center (www.predictioncenter.org) explained for the call for structural refinement competition. In addition, the ultimate goal in modeling protein structures is to understand their dynamics and functions. However, such detailed information is still very difficult to obtain through experiments directly, and hence developing theoretical methods can be very valuable and of great importance to assist studying these features of proteins.

In summary, the major challenges in the distance-based protein structure modeling are how to efficiently determine protein structures, or further refine protein structures and how to model protein dynamics, using the knowledge of interactions or inter-atomic distances between pairs of atoms. The primary motivation of my research is to investigate each of these problems in the field of the distance-based protein structure modeling. In particular, there are three main issues in my Ph.D. research work: i) solution of distance geometry problems, ii) protein structure refinement by the knowledge-based method, and iii) analysis of protein structural dynamics. Several algorithms and tools have been developed, which potentially have applications in related research fields. A brief introduction for each subject is provided.

Solution of distance geometry problems

The molecular distance geometry problem comes from the study of a molecular structure based on a set of inter-atomic distances; it has also an important application in structural biology, especially in protein structure prediction and determination [1]. In general, the distances between pairs of atoms can be obtained through physical experiments such as NMR experiments [2] or the knowledge of bond-lengths and bond-angles [3-4], or even knowledge-based methods such as structural alignment and homology modeling [5]. Then, the coordinates of atoms in a protein and hence the protein structure can be determined through solving distance-geometry problems, based on a set of inter-atomic distances. However, such problems have been proven to be NP-complete in general, and are especially difficult when only sparse and inexact distance data is available [4].

This subject was formally introduced by Blumenthal in 1953, who clearly explained that the distance geometry provides a way to find the coordinates of points in three-dimensional Euclidean space satisfying the given distances [6]. For the case in which all exact distances of a molecule are provided, the problem is relatively easy to solve. More specifically, it requires solving a singular value decomposition problem on distance matrix to find the coordinates of the points with the singular vectors. This problem is tractable and costs $O(n^3)$ floating point operations [7]. In practice, however only a sparse set of distances may be available. Then such a problem becomes very hard to solve and has been proved to be NP-hard by Saxe in 1979 [8]. In some other molecular applications, we can obtain lower and upper bounds on the distances. But such problems are still NP-hard as proved by More and Wu [9]. Traditional methods for solving distance geometry problems include singular value decomposition and the embedding algorithm by Crippen and Havel [7], the alternating projection algorithm by Glunt and Hayden [10], the graph reduction algorithm by Hendrickson [11],

the multi-scaling algorithm by Trosset [12], and the global smoothing algorithm by Moré and Wu [13-15] etc.

We have developed several algorithms for the solution of the distance geometry problem using a so-called geometric build-up approach, specifically for solving distance geometry problems with sparse but exact distances [16, 17]. In this approach, the coordinates of the atoms in a protein are determined one atom at a time, with the distances from four base atoms to the atom to be determined. In an ideal case, the coordinates of n atoms can then be determined in n steps, instead of n^2 steps as required by a conventional singular-value decomposition algorithm. However, a general geometric build-up algorithm can be numerically unstable for some cases when the numerical errors are accumulated in a long sequence of coordinate calculations. Also, the requirement for four base atoms for the unique determination of each atom is sufficient, but not necessary, and is even redundant for rigid determination.

We introduce the development of an updated geometric build-up algorithm that controls the increase of numerical errors [18] (see Chapter 2). The algorithm reinitializes the coordinates of the base atoms whenever necessary and possible, and can keep the errors from passing over to the atoms to be determined and resulting in incorrect structures. We also introduce a so-called rigid geometric build-up algorithm, which requires only three instead of four base atoms for the determination of each atom, and can generate rigid and sometimes, unique structures for very sparse distance data (see Chapter 3). The algorithm may produce multiple structures, due to the possible reflection for each atom. It keeps track of all combinations and in the end, determines a set of structures that are allowed by the given distances. We present the results obtained by using these algorithms for the determination or generation of the structures for a set of model proteins, and show the great potential of using the algorithms for protein structural analysis and determination.

Protein structure refinement by a knowledge-based method

Often, the protein structures determined by conventional experimental techniques usually are not as accurate as desired. Therefore, the usage of these protein structures usually has been severely limited in several important fields including homology modeling, drug design and protein dynamics. Further refinement is preferred and sometimes essential. As more and more structures are modeled and determined, the development of an efficient and reliable refinement technique becomes important, as the CASP prediction center in its call for structure refinement competition. In order to

refine the low resolution or low quality structures, many methods have been developed, including theoretical approaches [19-22] and knowledge based approaches [23-25].

Especially in recent years, with increasing numbers of high quality protein structures determined, the knowledge extracted from those proteins is a valuable source of information for protein structural analysis and structure determination. Considering the distance-based protein structure modeling, the knowledge of inter-atomic distances in proteins is also subject to certain statistics, and therefore obtaining additional distance information beyond the current theoretical and experimental limitations is very important and could be helpful to further protein structure refinement [25].

In this work, a computational approach for deriving mean-force potentials is developed for protein structure refinement, including constructing a database for protein inter-atomic distance distributions (PIDD) [26] (see Chapter 4). This database hosts and analyzes the statistical data for protein inter-atomic distances based on their distributions in databases of known protein structures such as in the Protein Data Bank (PDB). Further, we use the collected information to extract mean-force potentials which can be included in energy minimization, so the more plausible structural models may be determined (see Chapter 5).

We studied a set of NMR-determined protein structures by using the refinement approach with mean-force potentials. The improvements in the structures have been shown in terms of several standard measures, such as energy, RMSD and Ramachandran plots [27]. The method of mean force potentials has also been applied to comparative model refinement in the CASPR 2006 structural refinement (see Appendix A) (www.predictioncenter.org) and some important improvement has also been obtained. Together, these results imply that statistical information in distances is indeed valuable and could be applied to protein structure modeling.

Analysis of protein dynamics

The biological functions of proteins are highly correlated with their motions or flexibilities. General dynamic information and fluctuations can be always obtained experimentally in terms of B-factor and order parameters through Nuclear Magnetic Resonance (NMR) and X-ray Crystallography [28]. However, experimental analysis usually does not provide information much about the ways proteins move. Some theoretical methods such as all-atom molecular dynamics simulation have been applied to simulate protein dynamics [29], but all-atom simulation is very expensive in computation because of complicated potential energy functions. On the other hand, some simplified methods such

as Normal Mode Analysis (NMA) [30], Gaussian Network Model (GNM) [31] and Anisotropic Network Model (ANM) [32] have provided promising results comparable to those obtained by complicated methods that simulate protein dynamics. In general, such simplified methods involve fewer parameters and less detailed potential energy functions, and hence are more efficient in computation, compared to all-atom molecular dynamic simulations.

X-ray crystallography determines a unique protein structure with high resolution and quality, while NMR determines an ensemble of multiple energy-minimized structures satisfying distance constraints, rather than a unique conformation. Sometimes, there is significant difference between models in an ensemble [25, 33]. In comparison with crystal structures, there are not many sophisticated methods developed to theoretically study fluctuations and dynamics of NMR-determined ensembles.

Here we investigate a new computational approach to study protein dynamics of NMR ensembles at the residue level (only Cα atoms). In this work, we modified distance matrix error (DME) calculations to be locally specific. For each Cα atom, only distances between it and other atoms are considered, and differences of those distances between all possible pairs in two structures in an NMR ensemble are summed and represent its flexibility. We compared the Local-DME values of NMR-determined proteins with B factor values for the same proteins determined by X-ray crystallography. The High correlation obtained indicates the possibility of using Local-DME calculations to compute pseudo B factor values of NMR ensembles and to provide an alternative way for investigating protein dynamics in solution.

Organization of thesis

The thesis is organized as follows. In Chapter 2, we introduce the geometric build-up algorithm and the numerical problems existing in this algorithm. We then describe the updated geometric build-up algorithm and discuss related numerical issues. Some numerical results obtained by applying the updated algorithm are presented and compared with the general algorithm. I am the major contributor to this paper and Dr Zhijun Wu provided me very suggestive comments. In Chapter 3, we describe a rigid geometric-up algorithm for very sparse distance data, and also try to investigate the sufficient and necessary condition of protein structure determination. The motivation and related numerical issues are addressed. And conclusion remarks have also been based on the numerical testing on a set of proteins with sparse distance data are included. I am the major contributor to this paper. In Chapter 4, a description of our database for protein inter-atomic distance distributions is

presented. We also provide the architecture of this database and related research using this database, such as generating additional distance constraints and deriving distance-based potentials. I setup the database, wrote the interface and now maintain the database. In Chapter 5, we describe a novel knowledge-based method for protein structure refinement. We provide a systematic introduction to NMR protein structure determination as well as the current challenges. A detailed refinement protocol using potentials derived from distance distributions is discussed. Testing results on 70 NMR-determined structures are also been shown. I was conducting the entire work with the help from Dr. Wu and Dr Jernigan. In Chapter 6, we introduce the study of protein structure dynamics as well as current computational approaches. We review the Gaussian Network Model and introduce an efficient and reliable computational tool, called Local-DME calculation, for studying protein dynamics of NMR ensembles in solution. Comparison of Local-DME values, experimental B factor values and fluctuations predicted by GNM is also provided for a test set of protein structures. I was conducting the entire work with the help from Dr. Wu and Dr Jernigan. In Chapter 7, the entire thesis work is summarized and some important issues for future investigation are discussed. In Appendix A, the source code in Matlab used in geometric build-up algorithms is shown. In Appendix B, the source code of the interface of PIDD database is shown. In Appendix C, we displayed part of the tutorial of PIDD database. In Appendix D, the FORTRAN source code of the subroutine for the mean force potential is provided. In Appendix E, we will discuss the work in refining comparative models using mean force potentials, and especially the testing results of a target structure in the CASPR competition are investigated and analyzed. In Appendix F, the Matlab code for calculating the Local-DME values is provided. All appendixes are completed by me.

References

1. Yoon J, Gad Y, and Wu Z, Mathematical Modeling of Protein Structure with Distance Geometry, *Numerical Linear Algebra and Optimization*, Scientific Press, 2002.
2. Wuthrich, K., NMR of Proteins and Nucleic Acids, Wiley, New York, 1986
3. Brooks C, Karplus M, and Pettitt B, Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, John Wiley & Sons, 1988.
4. Creighton T, Proteins: Structures and Molecular Properties, 2nd Edition, Freeman and Company, 1993.
5. Havel T, Snow M. 1991. A new method for building protein conformations from sequence alignments with homologues of known structure. *Journal of Molecular Biology*, 217:1-7.

6. Blumenthal L, Theorey and Applications of Distance Geometry, Oxford, Clarendon Press, 1953
7. Crippen G and Havel T, Distance Geometry and Molecular Conformation, John Wiley & Sons, 1988.
8. Saxe J, Embeddability of Weighted Graphs in K-Space Is Strongly NP-Hard, *in Proc. 17th Allerton Conference in Communications, Control and Computing*, 1979, pp. 480-489.
9. Moré J and Wu Z, ε-Optimal Solutions to Distance Geometry Problems via Global Continuation, in Global Minimization of Non-Convex Energy Functions: Molecular Conformation and Protein Folding, American Mathematical Society, 1996a, pp. 151-168.
10. Glunt W, Hayden T, Hong S, and Wells J, An Alternating Projection Algorithm for Computing the Nearest Euclidean Distance Matrix, *SIAM Journal of Mathematical Analysis and Applications,* Vol. 11, No 4, 1990, pp. 589-600.
11. Hendrickson B, The Molecular Problem: Determining Conformation from Pairwise Distances, Ph.D. thesis, Cornell University, 1991.
12. Trosset, M, Applications of Multidimensional Scaling to Molecular Conformation, *Computing Sciences and Statistics*, 29, 1998, pp. 148-152.
13. Moré J and Wu Z, Smoothing Techniques for Macromolecular Global Optimization, in Nonlinear Optimization and Applications, Plenum Press, 1996b, pp. 297-312.
14. Moré J and Wu Z, Global Continuation for Distance Geometry Problems, *SIAM Journal of Optimization*, Vol. 7, No. 3, 1997a, pp. 814-836.
15. Moré J and Wu Z, Issues in Large Scale Global Molecular Optimization, in Large Scale Optimization with Applications, Springer-Verlag, 1997b, pp. 99-122.
16. Dong Q and Wu Z, A Linear-Time Algorithm for Solving the Molecular Distance Geometry Problem with Exact Inter-Atomic Distances, *Journal of Global Optimization*, Vol. 22, 2002, pp. 365-375.
17. Dong Q and Wu Z, A Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse Distance Data, *Journal of Global Optimization*, Vol. 26, 2003, pp. 321-333.
18. Wu D, and Wu Z, An Updated Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse Distance Data. *Journal of Global Optimization*, 2006 (accepted).
19. Clore G, Gronenborn A, New methods of structure refinement for macromolecular structure determination by NMR. *Proceedings of the National Academy of Sciences,* 95. 5891-5898 (1998).

20. Chen J, Im W, Brooks CL 3rd. Refinement of NMR structures using implicit solvent and advanced sampling techniques. *Journal of American Chemistry Society*. 2004 Dec 15;126(49):16038-47.
21. Xia B, Tsui V, Case D, Dyson H, Wright P. Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water. *Journal of Biomolecular NMR*. 2002 Apr;22(4):317-31.
22. Linge J, Nilges M. Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. *Journal of Biomolecular NMR*. 1999 Jan;13(1):51-9.
23. Kuszewski J, Gronenborn A, and Clore G. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Science* 5. 1067-1080 (1996).
24. Grishaev A. and Bax A. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. *Journal of American Chemistry Society*. 126. 7281-7292 (2004).
25. Cui F, Jernigan R, Wu Zj, Refinement of NMR-determined protein structures with database derived distance constraints. *Journal of Bioinformatics and Computational Biology*, 2005, 3(6):1315-29.
26. Wu D, Cui F, Jernigan R, and Wu Z, PIDD: A database for protein inter-atomic distance distribution, submitted, 2006
27. Ramachandran G and Sasiskharan V, Conformation of polypeptides and proteins. *Advanced Protein Chemistry*, 1968,23:283-437
28. Karplus M, McCammon J. The internal dynamics of globular proteins. *Critical Reviews in Biochemistry and Molecular Biology*. 1981;9(4):293–349.
29. McCammon J, Wolynes P, Karplus M. Picosecond dynamics of tyrosine side chains in proteins. *Biochemistry*. 1979 Mar 20;18(6):927-42.
30. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of Molecular Biololgy*. 1985 Feb 5;181(3):423-47.
31. Haliloglu T, Bahar I, Erman, Gaussian dynamics of folded proteins.,B. *Physics Review Letter* 79, 3090-3093, 1997.
32. Atilgan R, Durell S, Jernigan R, Demirel M, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysics Journal*, 2001 80:505-515.

33. Zhao D, Jardetzky O. An assessment of the precision and accuracy of protein structures determined by NMR. Dependence on distance errors. *Journal of Molecular Biology,* 1994 Jun 24;239(5):601-7.

CHAPTER 2. AN UPDATED GEOMETRIC BUILD-UP ALGORITHM

A paper accepted by the Journal of Global Optimization with the complete name an updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distance data

Di Wu and Zhijun Wu

Abstract

An updated geometric build-up algorithm is developed for solving the molecular distance geometry problem with a sparse set of inter-atomic distances. Different from the general geometric build-up algorithm, the updated algorithm re-computes the coordinates of the base atoms whenever necessary and possible. In this way, the errors introduced in solving the algebraic equations for the determination of the coordinates of the atoms are controlled in the intermediate computational steps. The method for re-computing the coordinates of the base atoms based on the estimation on the root-mean-square deviation is described. The results of applying the updated algorithm to a set of protein structure problems are presented. In many cases, the updated algorithm solves the problems with high accuracy when the results of the general algorithm are inadequate.

Keywords Protein structure determination, distance geometry, geometric build-up, root-mean-square deviation

Introduction

The molecular distance geometry problem arises in the study of the structure of a molecule based on a given set of inter-atomic distances for the molecule. This problem has an important application in molecular biology and biochemistry, and in particular, in protein structure prediction and determination (see Yoon, Gad, and Wu 2002 for a general review). The distances between certain pairs of atoms in protein can often be determined based on our knowledge of various types of bond-

lengths and bond-angles (Brooks III, Karplus, and Pettitt 1988, Creighton 1993), or from nuclear magnetic resonance (NMR) experiments (Brüger and Niles 1993, Kuntz, Thomason, and Oshiro 1993), or sometimes, through homology modeling (Havel and Snow 1991). Therefore, a natural approach for the determination of the structure of a protein is to solve a molecular distance geometry problem if a set of distance data for the protein is given. However, the molecular distance geometry problem is difficult to solve in general, especially since often in practice, only sparse and inexact distance data is available. Several algorithms have been developed to solve the problem, including for example the embed algorithm by Crippen and Havel (1988), the alternating projection algorithm by Glunt and Hayden (1990, 1993), the graph reduction algorithm by Hendrickson (1991, 1995), the multi-scaling algorithm by Trosset (1997) and Kearsly, Tapia, and Trosset (1998), the global smoothing algorithm by Moré and Wu (1996a, 1996b, 1997a, 1997b, 1999), etc. Most of these algorithms can provide an approximate solution to the problem, but often not to a desired accuracy. They are costly requiring intensive computation as well.

In their recent work, Dong and Wu (2002a) proposed a new approach to the molecular distance geometry problem. This approach, called the geometric build-up approach, determines the coordinates of the atoms in the molecule one atom at a time repeatedly using a simple geometric relationship between determined and undetermined atoms, i.e., if an undetermined atom has known distances to four previously determined atoms and if the four atoms are not in the same plane, then it is a simple geometric fact that the coordinates of the undetermined atom can immediately be determined by using the four known distances (see also Huang, Liang, and Pardalos 2002 for more general discussions on these properties). If the exact distances between all pairs of atoms are given, this approach can determine the coordinates of n atoms in n steps or in other words, in order of n floating point operations, while a conventional singular-value decomposition algorithm (as used in the embed algorithm) requires at least order of n2 floating point operations.

In this paper, we consider the solution of a molecular distance geometry problem with sparse but exact distance data by using a geometric build-up algorithm. For such a problem, since the data is sparse, the required distances may not be available when an atom is to be determined. The atom is then put aside until the distances become available after more atoms are determined. For this purpose, the algorithm is applied repeatedly to the undetermined atoms until all remaining ones are determined. Dong and Wu (2002b) implemented such an algorithm, but they found that the algorithm is very sensitive to the numerical errors introduced in calculating the coordinates of the atoms. The reason is that the coordinates of the atoms are all determined using the coordinates of previously determined atoms, and the errors in the previously determined atoms are passed to and accumulated in later

determined atoms. As a result, the coordinates for later determined atoms become incorrect, especially when the molecule is large, say with more than a thousand atoms. Note that this problem does not exist for the problem with all exact distances since in that case we can just use one set of determined atoms to determine all other atoms and there will not be a chance for the errors to get propagated.

In this paper, we describe a so-called updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse but exact distance data. We show that using this algorithm the accumulation of the errors in calculating the coordinates of the atoms can be controlled and prevented. The idea for the algorithm is based on the fact that the coordinates of any four atoms can be determined without any other information as long as all distances among them are given. For this reason, the coordinates of any four determined atoms can be re-calculated whenever possible using the distances among them if the distances are given. The re-calculated coordinates do not depend on the coordinates of previously determined atoms and therefore do not inherit any errors from them. In this way, the coordinates for many of the atoms can be "corrected", and the errors in the calculated coordinates can be prevented from growing into incorrect structural results. The re-calculated coordinates for the four atoms are independent of their original coordinates and are not related to the overall structure already built-up by the algorithm. However, they can be put back to the original structure by aligning them to their original locations with an appropriate translation and rotation.

The general geometric build-up algorithm

A geometric build-up algorithm for solving the molecular distance geometry problem given the exact distances between all pairs of atoms in the molecule is outlined in Figure 1. There are two parts in the algorithm. The first one is to select four initial atoms that are not in the same plane and find a set of coordinates for the atoms using the distances among them. Let us call the atoms the base atoms. After a set of base atoms is selected and allocated, the second part of the algorithm is to find the coordinates for each of the remaining atoms using the distances from the atoms to the four base ones. The first part of the algorithm is based on the fact that the coordinates of four atoms can be determined if all distances among them are given, while the second part is that the coordinates of an atom can be determined if the distances from the atom to four determined atoms are given. In both cases, the coordinates can be determined through simple algebraic calculations and in particular, for

the latter case, through the solution of a small system of algebraic equations. We state these facts in a more rigorous form in the following theorems.

Figure 1. The outline of the geometric build-up algorithm

The Geometric Build-up Algorithm for Problems with All Exact Distances*

1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. For each of the remaining atoms,
   determine the coordinates of the atom with its distances to the base atoms.
3. All atoms are determined.

*The outline of the general geometric build-up algorithm for solving the molecular distance geometry problem with all exact distances (Dong and Wu 2002a)

Theorem 2.1. If the distances among four atoms are given, the coordinates of the atoms can then be determined with the given distances, subject to translation, rotation, and reflection.

Proof. Let $x_i = (u_i, v_i, w_i)^T$, $i = 1, 2, 3, 4$, be the coordinate vectors of the four atoms. Let $d_{i,j}$ be the given distances between atoms i and j for $i, j = 1, 2, 3, 4$. The coordinates can then be determined as follows, based on the given distances.

First, since the atoms can be allocated in an arbitrary coordinate system, without loss of generality, we set a system with the first atom at its origin, the second on its x-axis, and the third on its xy-plane. Then, we have in this system that $u_1=0$, $v_1=0$, $w_1=0$, $v_2=0$, $w_2=0$, and $w_3=0$. Since the distance from the second atom to the first atom is equal to $d_{2,1}$, we have also that $u_2=d_{2,1}$, and the first two atoms are then determined.

Since the distances from the third atom to the first and second atoms are equal to $d_{3,1}$ and $d_{3,2}$, respectively, then

$$u_3^2 + v_3^2 = d_{3,1}^2$$
$$(u_3 - u_2)^2 + v_3^2 = d_{3,2}^2.$$

Solve the equations for u_3 and v_3. We obtain

$$u_3 = (d_{3,1}^2 - d_{3,2}^2)/(2u_2) + u_2/2$$
$$v_3 = \pm(d_{3,1}^2 - u_3^2)^{1/2},$$

and the third atom is then determined by choosing v_3 either positive or negative.

Finally, with the distances, $d_{4,1}$, $d_{4,2}$, $d_{4,3}$, from the fourth atom to the first three atoms, we can form three equations,

$$u_4^2 + v_4^2 + w_4^2 = d_{4,1}^2$$
$$(u_4 - u_2)^2 + v_4^2 + w_4^2 = d_{4,2}^2$$
$$(u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2 = d_{4,3}^2.$$

The coordinates u_4, v_4, w_4 for the fourth atom can then be determined by solving the equations, and

$$u_4 = (d_{4,1}^2 - d_{4,2}^2)/(2u_2) + u_2/2$$
$$v_4 = (d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2)/(2v_3) + v_3/2$$
$$w_4 = \pm(d_{4,1}^2 - u_4^2 - v_4^2)^{1/2}.$$

This completes the proof for Theorem 2.1.

Theorem 2.2. If the coordinates of four atoms that are not in the same plane and the distances from the fifth atom to the four atoms are given, the coordinates of the fifth atom can be determined uniquely.

Proof. Let $x_i = (u_i, v_i, w_i)^T$, $i = 1, 2, 3, 4$, be the coordinate vectors of the first four atoms and $x_j = (u_j, v_j, w_j)^T$ the coordinate vector of the fifth atom with an arbitrary index j. Let $d_{i,j}$ be the given distances from any of the first four atoms i to the fifth atom j for $i = 1, 2, 3, 4$. We then have a set of equations,

$$\| x_i - x_j \| = d_{i,j}, \quad i = 1,2,3,4.$$

Square the equations and expand their left-hand-sides to obtain

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1,2,3,4.$$

Subtract the first equation from the rest to reduce the equations to the following three,

$$-2(x_{i+1} - x_i)^T x_j = (d_{i+1,j}^2 - d_{i,j}^2) - (\| x_{i+1} \|^2 - \|x_i\|^2), \quad i = 1,2,3.$$

Let A be a matrix and b a vector, and

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_1)^T \\ (x_4 - x_1)^T \end{bmatrix},$$

$$b = \begin{bmatrix} (d_{2,j}^2 - d_{1,j}^2) - (\| x_2 \|^2 - \|x_1\|^2) \\ (d_{3,j}^2 - d_{1,j}^2) - (\| x_3 \|^2 - \|x_1\|^2) \\ (d_{4,j}^2 - d_{1,j}^2) - (\| x_4 \|^2 - \|x_1\|^2) \end{bmatrix}.$$

We can then write the above equations in the following matrix form.

$$A x_j = b$$

Since x_1, x_2, x_3, x_4 are not in the same plane, the matrix A is nonsingular and therefore, the linear system of equations can be solved to obtain a unique solution for x_j.

Note that Theorems 2.1 and 2.2 both assume that the given distances are accurate and consistent, and are true distances among a set of points. Given such distances, the coordinates of the atoms can obviously be determined by using the algorithm described in Figure 1 based on the two theorems. Moreover, it can be proved that the coordinates of the atoms for a molecule of n atoms can be determined in n steps, each for one atom, as stated in the following theorem.

Theorem 2.3. The general geometric build-up algorithm solves a molecular distance geometry problem with all exact distances for a molecule of n atoms in order of n floating point operations or in other words, in linear time in n.

Proof. As shown in Figure 1, once the base atoms are determined, the remaining atoms are determined using the distances from the atoms to the base atoms, each requiring the solution of a small linear system of equations based on Theorem 2.2. Solving the linear system can be done in constant time, so for all remaining atoms, the time for determining them all is proportional to the number of atoms, n-4. The determination of the coordinates of the base atoms does not cost more than constant time, but to make sure the base atoms are not in the same plane may take longer time. In the worst case, the latter may take order of n computing time to examine through the entire atom list to find the third atom that is not in the line formed by the first two atoms ($v_3 \neq 0$) and then the fourth atom that is not in the plane formed by the first three atoms ($w_4 \neq 0$). In any case, the algorithm requires order of n floating-point operations or in other words, linear time in n to find the coordinates of all n atoms.

We now consider the case when only a subset of all distances among the atoms is available. The problem can be called one with sparse exact distances. In this case, the algorithm in Figure 1 will not work since the required distances from the base atoms to the atom to be determined may not be available. However, the distances from other determined atoms to the atom may be available and may

suffice for the determination of the atom. Therefore, the algorithm can be modified to cover the sparse case by determining the coordinates of an atom using any determined atoms as long as they can serve as its base atoms. Such a modified algorithm is outlined in Figure 2.

Figure 2. The outline of geometric build-up algorithm for sparse data

The Geometric Build-Up Algorithm for Problems with Sparse Exact Distances*

1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. Repeat:
   For each of the remaining atoms,
      find four determined atoms that can serve as its base atoms;
      determine the coordinates of the atom with its distances to the base atoms.
   End
   If no atom is determined in the whole loop, stop.
3. All atoms are determined.

*The outline of the general geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distances (Dong and Wu 2002b)

Note that when only a sparse set of distances is given, the molecular distance geometry problem becomes difficult to solve in general. We do not expect to have a polynomial time algorithm for the problem, since Saxe (1979) has proved that the problem actually becomes *NP*-complete. Also, in the algorithm outlined in Figure 2, the four qualified base atoms may not be available in the first step anyway; the for-loop in the second step may be repeated many times until all remaining atoms can be determined. However, Dong and Wu (2002b) demonstrated that for protein structure determination, the algorithm seemed to be a reasonable one. When the distances less than 8 Å were used, reasonable structures for a set of tested proteins with up to 4200 atoms were obtained by using such an algorithm. A numerical problem in this algorithm, as pointed out in Dong and Wu (2002b), is that the base atoms that are used to determine an atom are determined themselves by some other base atoms in previous steps. The errors introduced in previous steps are thus passed to the current atom, and to the atoms in later steps as well. This may cause a completely incorrect result in the coordinates of the atoms. The errors in calculating the coordinates of an atom usually come from solving the linear system of equations, especially if the coefficient matrix A is ill formed. The matrix A is determined by the coordinates of the base atoms as shown in the proof for Theorem 2.2. Therefore, in

Dong and Wu (2002b), if the determinant of A is found small, a different set of base atoms would be used to avoid possible errors due to this matrix A, which resolved the problem for some of the test cases, but not for all.

The updated geometric build-up algorithm

In this section we describe the updated geometric build-up algorithm. The algorithm is a modified version of the general algorithm for problems with sparse exact distances. Two new strategies are used to minimize the errors introduced in the coordinate calculations. First, the condition number instead of the determinant of matrix A is examined when solving each of the linear systems in the algorithm. When the condition number is too big, a different set of base atoms is sought to avoid the possible errors due to an ill-conditioned matrix A. This is better than evaluating the determinant since a matrix can still be ill conditioned even if its determinant is large. Second, the coordinates of four determined atoms are re-calculated or re-initialized by the procedure described in Theorem 2.1, whenever the four atoms are found that they have all distances available among them. Since they are independent of the coordinates of previously determined atoms, the re-calculated coordinates do not have the errors accumulated from previous calculations and hence re-calculation of coordinates reduces the chance of error accumulation. As described in the proof for Theorem 2.1, the re-calculated coordinates are represented in a new coordinate system with one atom located in the origin, another along the x-axis, etc. However, the atoms can be put back to the original structure by aligning their new coordinates with the old ones, using an appropriate translation and rotation for the new coordinates, so that the RMSD between the new coordinates and the old ones is minimized. The translation vector and the rotation matrix can be obtained exactly in the same way as in regular RMSD calculations.

Figure 3 is an outline of the updated algorithm. We call it updated since the coordinates are updated repeatedly in the algorithm to prevent errors. The way we calculate the RMSD of two structures (defined in terms of their Cartesian coordinates) is the following. Let X and Y be the coordinate matrices of two structures after they are translated so that their centers of geometry coincide. The RMSD of the two structures is then defined as

$$\mathrm{RMSD}(X,Y) = \min_Q \| X - YQ \|_F / \sqrt{n},$$

where Q is a rotation matrix and $QQ^T = I$. Let $C = Y^TX$, and let $C = U\Sigma V^T$ be the singular-value decomposition of C. Then it is not difficult to verify that $Q=UV^T$ solves the above minimization problem (Golub and van Loan 1989).

Figure 3. The outline of an updated geometric build-up algorithm

The Updated Geometric Build-Up Algorithm for Problems with Sparse Exact Distances*

1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. Repeat:
   For each of the remaining atoms,
   find four determined atoms that can serve as its base atoms;
   determine the coordinates of the atom with its distances to the base atoms.
   If four determined atoms are found having all distances among them,
   re-initialize the coordinates of the four atoms;
   put the atoms back to the original structure.
   End
   End
   If no atom is determined in the whole loop, stop.
3. All atoms are determined.

*The outline of the updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distances.

Therefore, computationally, we can first compute the geometric centers of the two structures,

$$xc = \frac{1}{n}\sum_{i=1}^{n} X(i,:), \quad yc = \frac{1}{n}\sum_{i=1}^{n} Y(i,:).$$

We then update matrix Y,

$$Y(:,1) = Y(:,1) - [yc(1) - xc(1)],$$
$$Y(:,2) = Y(:,2) - [yc(2) - xc(2)],$$
$$Y(:,3) = Y(:,3) - [yc(3) - xc(3)].$$

The two structures now have the same geometric center. We then compute the matrix $C=Y^TX$ and its singular-value decomposition $C=U\Sigma V^T$. Let $Q=UV^T$. The RMSD of the two structures can then be calculated as

$$\mathrm{RMSD}(X,Y) = \| X - YQ \|_F / \sqrt{n}.$$

In the updated algorithm, every time the coordinates of four atoms are re-calculated, if X contains the old coordinates and Y the new ones, YQ in the above formula gives the coordinates best aligned with the old ones.

Numerical results

We have implemented an updated geometric build-up algorithm in Matlab (Version 5.3) (see the source code in Appendix A). The matrix-vector calculations required in the algorithm including linear system solves, estimations of condition numbers, and singular-value decompositions are all done through the Matlab build-in functions. We tested the algorithm with a set of problems generated using the known structures of ten proteins downloaded from the PDB data bank (Berman et al, 2000). Each of the structures is used to obtain two sets of distances, one including all distances ≤ 5 Å and another ≤ 8 Å. We then solve a molecular distance geometry problem for each set of distances using the updated algorithm, to obtain the coordinates of the atoms for the corresponding protein. The result is compared with the original structure of the protein in terms of RMSD. The choice of 5 Å as the cut-off distance is made to simulate the distance data in NMR experiments since in most cases, NMR can only detect the distances between atoms in that range. The choice of 8 Å is to make a relaxation on the cut-off to observe the performance difference of the algorithm under a different condition. Note that in practice, NMR actually can provide only lower and upper bounds of the distances. However, in this work, we only consider problems with exact distances. The extension of the algorithm to problems with distance bounds is possible and under another line of investigation.

Table 1 contains the results of using the updated geometric build-up algorithm for solving the generated test problems. They are also compared with the results of using the general geometric build-up algorithm for the same set of problems obtained by Dong and Wu (2002b). The first column of the table contains the names of the proteins in the PDB Data Bank. The second column contains the numbers of atoms in the proteins. The remaining columns list the results of using the updated and general algorithms for problems with 5 Å and 8 Å distance cut-offs. The results for each problem include the number of fixed atoms and the RMSD for the fixed structure compared with the original one. For the ten tested structures, five of them were determined with the general algorithm with the distances less than 8 Å, but none with less than 5 Å. However, nine of the ten structures were determined with the updated algorithm with the distances less than 8 Å, and five of them were determined with the distances less than 5 Å. For the updated algorithm, we also list the results for problems that were not completely resolved by the algorithm with the distances less than 5 Å. They include 1PHT, 1AX8, 1RGS, 1BPM, and 1HMV. These problems are relatively large, but for four of them, the algorithm actually was able to determine the coordinates for almost all the atoms. For 1PHT only 5 out of 814 atoms were not fixed, and for 1RGS only 5 out of 2015, for 1BPM only 3 out of 3674, and for 1HMV only 13 out of 4201. We have examined the atoms that were not fixed by the

algorithm and found that in many cases, the atoms are in the side chains of the proteins and do not have enough neighboring atoms within 5 Å distance. For example, in 1PHT, the unfixed atoms are located in the side chain of LYS, where there are not enough distances to determine the atoms. The structure 1AX8 seems difficult to determine probably because it is a double helix and is lack of enough distance information among the atoms. We include this instance in the table to show the possible difficult case for the algorithm. There are two odds in the table requiring some explanations as well. First, for 1PTQ, with an 8 Å cut-off, the number of fixed atoms for the general algorithm is 402 instead of 404. This is because that there are two heat atoms in the structure that were not considered in the experiment with the general algorithm. Second, for 1AX8, with an 8 Å cut-off, 988 out of 1003 atoms were determined using the updated algorithm, but all atoms were determined using the general algorithm. This may be because of the specific structure of the molecule or the specific implementation of the two algorithms, but it does not reflect the general behaviors of the algorithms.

Table 1. Results of the updated geometric build-up algorithm

Protein*	#atom	5 Å				8 Å			
		Updated		General		Updated		General	
		#fixed atom	RMSD	#fixed atom	RMSD	#fixed atom	RMSD	#fixed atom	RMSD
1PTQ	404	404	2.7e-012	--	--	404	3.5e-013	402	2.8e-008
1HOE	558	558	8.2e-013	--	--	558	1.0e-011	558	9.4e-006
1LFB	641	641	9.5e-012	--	--	641	3.9e-012	--	--
1F39A	767	767	3.5e-011	--	--	767	2.4e-012	767	2.3e-006
1PHT	814	809	7.9e-009	--	--	814	1.8e-012	814	4.4e-005
1POA	914	914	6.8e-010	--	--	914	1.7e-011	--	--
1AX8	1003	--	--	--	--	998	3.5e-012	1003	1.5e-006
1RGS	2015	2010	7.4e-008	--	--	2015	1.1e-009	--	--
1BPM	3674	3671	1.8e-009	--	--	3674	3.2e-007	--	--
1HMV	4201	4188	6.8e-011	--	--	4201	2.5e-005	--	--

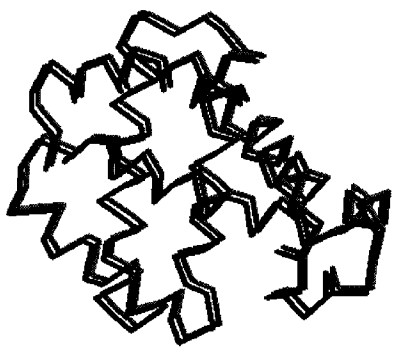
* Results of using the updated and general geometric build-up algorithms for solving a set of molecular distance geometry problems generated from ten known protein structures downloaded from the PDB Data Bank.

Figures 4 and 5 further demonstrate in some worst-case scenarios how the structure determined by a geometric build-up algorithm can be affected by the accumulated numerical errors. The figures show the structures (red lines) of protein 4MBA (1086 atoms) determined using ≤ 5 Å distances, first by a general algorithm and then by the updated algorithm. The pictures show clearly that the general algorithm results in a structure (red lines in Figure 4) that disagrees with the original structure (blue lines) in many regions, while the updated algorithm determines one (red lines in Figure 5) that agrees with the original structure (blue lines) almost completely.

Figure 4. The structure of 4MBA by the general algorithm*

*The structure (red lines) of 4MBA determined by using a general geometric build-up algorithm and compared with the original structure of 4MBA (blue lines). Here, 4MBA is the PDB entry for the crystal structure of the ferric form of myoglobin from the mollusc Aplysia limacina refined at 1.6 Å resolution, by restrained crystallographic refinement methods. The crystallographic R-factor is 0.19. The tertiary structure of the molecule conforms to the common globin fold, consisting of eight alpha-helices. The N-terminal helix A and helix G deviate significantly from linearity. See Bolognesi et al, (1989) for more details.

Figure 5. The structure of 4MBA generated by the updated algorithm*

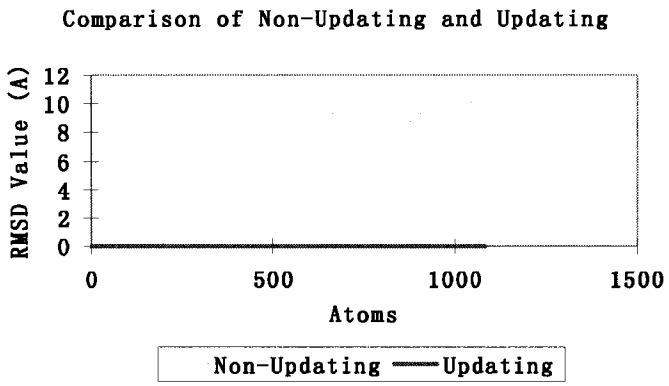


*The structure (red lines) of 4MBA determined by using an updated geometric build-up algorithm and compared with the original structure of 4MBA (blue lines)

Finally, Figure 6 further shows how the numerical error grows as the geometric build-up algorithm proceeds. Shown in the figure is the RMSD of the computed structure for 4MBA compared with its original structure as a function of the size (the number of atoms) of the computed structure. For a general geometric build-up algorithm, from around 300 atoms, the RMSD (the green line) starts increasing rapidly, and in the end, the RMSD for the entire structure (with 1086 atoms) becomes

bigger than 10 Å. On the other hand, for the updated algorithm, the RMSD (the blue line) is bounded in around 5.0e-04 Å in the whole build-up procedure.

Figure 6. Numerical errors by the updated and general algorithms



Summary and remarks

The molecular distance geometry problem has an important application in macromolecular modeling and in particular, in protein structure determination. The problem is difficult to solve especially in practice when only sparse and inexact distances are given. In this paper, we consider the solution of a molecular distance geometry problem with sparse but exact distance data by using a geometric build-up algorithm. For such a problem, since the data is sparse, the coordinates of the atoms cannot be determined with only one set of base atoms since the required distances between the base atoms and the atom to be determined may not be available. Therefore, in most cases, the atoms are determined using a set of base atoms that are determined in previous steps. Dong and Wu (2002b) implemented such an algorithm, but they found that the algorithm is very sensitive to the numerical errors introduced in calculating the coordinates of the atoms. The reason is that the coordinates of the atoms depend on the coordinates of previously determined atoms, and the errors in the previously determined atoms are passed to and accumulated in later determined atoms. As a result, the coordinates for later determined atoms become incorrect, especially when the molecule is large, say with more than a thousand atoms.

In this paper, we have introduced an updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse but exact distance data. We have shown that using this algorithm the accumulation of the errors in calculating the coordinates of the atoms could be controlled and prevented. The idea for the updated algorithm is based on the fact that the coordinates

of any four atoms can be determined without any other information as long as all distances among them are given. Therefore, the coordinates of any four determined atoms can be re-calculated whenever possible using the distances among them if the distances are given. The re-calculated coordinates do not depend on the coordinates of previously determined atoms and therefore do not inherit any errors from them. In this way, the coordinates for many of the atoms can be "corrected", and the errors in the calculated coordinates can be prevented from growing into incorrect structural results.

We have described the general geometric build-up algorithm with a presentation that is more formal than that of other papers. Several important properties related to the algorithm are stated as theorems and formal proofs are also given. Some of them are the foundations for the development of the general as well as updated geometric build-up algorithms. We have discussed the numerical issues associated with the general geometric build-up algorithm and presented the updated algorithm including the procedure for re-evaluating the coordinates and the method for updating the old coordinates with the new ones through RMSD calculation. We have presented numerical results of using the updated algorithm for a set of test problems generated with known protein structures. The results for two sets of problems have been obtained, one with distances less than or equal to 5 Å and another 8 Å. The results showed that the updated algorithm determined the structures for most of the problems while the general algorithm failed.

The algorithm discussed in this paper may be of only theoretical value in a certain sense since in practice the given distances usually are inexact and the algorithm may only be used for solving a sub-problem. However, the algorithm represents a significant advance in solving a general molecular distance geometry problem. It can certainly be modified and extended to problems with inexact distances. Work in this direction is being pursued and will be reported later elsewhere.

Acknowledgements

We would like to thank Peter Vedell for reading the paper and offering helpful suggestions. The support for the first author from the ISU Graduate Program on Bioinformatics and Computational Biology is also gratefully acknowledged.

References

1. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, L. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, *Nuc. Acid. Res.,* Vol 28, 2000, pp. 235-242.
2. M. Bolognesi, S. Onesti, G. Gatti, A. Coda, P. Ascenzi, and M. Brunori, Aplysia Limacina Myoglobin: Crystallographic Analysis at 1.6 Å Resolution. *J. Mol. Biol.*, Vol. 205, 1989, pp. 529-544.
3. C. L. Brooks III, M. Karplus, and B. M. Pettitt, Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, John Wiley & Sons, 1988.
4. A.T. Brüger and M. Niles, Computational Challenges for Macromolecular Modeling, in Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publishers, 1993, Vol. 5, pp. 299-335.
5. T. E. Creighton, Proteins: Structures and Molecular Properties, 2nd Edition, Freeman and Company, 1993.
6. G. M. Crippen and T. F. Havel, Distance Geometry and Molecular Conformation, John Wiley & Sons, 1988.
7. Q. Dong and Z. Wu, A Linear-Time Algorithm for Solving the Molecular Distance Geometry Problem with Exact Inter-Atomic Distances, *J. Global Optim.,* Vol. 22, 2002, pp. 365-375.
8. Q. Dong and Z. Wu, A Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse Distance Data, *J. Global. Optim.,* Vol. 26, 2003, pp. 321-333.
9. W. Glunt, T. L. Hayden, S. Hong, and J. Wells, An Alternating Projection Algorithm for Computing the Nearest Euclidean Distance Matrix, *SIAM J. Mat. Anal. Appl.,* Vol. 11, No 4, 1990, pp. 589-600.
10. W. Glunt and T. L. Hayden and M. Raydan, Molecular Conformations from Distance Matrices, *J. Comput. Chem.,* Vol. 14, No. 1, pp. 114-120, 1993.
11. G. H. Golub and C. F. van Loan, Matrix Computations, Johns Hopkins University Press, 1989.
12. T. F. Havel, Distance Geometry, in Encyclopedia of Nuclear Magnetic Resonance, D. M. Grant and R. K. Harris, eds., John Wiley & Sons, 1995, pp. 1701-1710.
13. T. F. Havel and M. E. Snow, A New Method for Building Protein Conformations from Sequence Alignments with Homologues of Known Structure, *J. Mol. Biol.,* Vol. 217, 1991, pp. 1-7.
14. B. A. Hendrickson, The Molecular Problem: Determining Conformation from Pairwise Distances, Ph.D. thesis, Cornell University, 1991.

15. B. A. Hendrickson, The molecule problem: Exploiting Structure in Global Optimization, *SIAM J. Optim.*, Vol. 5, No. 4, 1995, pp. 835-857.
16. H. X. Huang and Z. A. Liang, and P. Pardalos, Some Properties for the Euclidean Distance Matrix and Positive Semi-Definite Matrix Completion Problems, Department of Industrial and Systems Engineering, University of Florida, 2002.
17. A. Kearsly, R. Tapia, and M. Trosset, Solution of the Metric STRESS and SSTRESS Problems in Multidimensional Scaling by Newton's Method, *Computational Statistics* 13, 1998, pp. 369-396.
18. D. Kuntz, J. F. Thomason, and C. M. Oshiro, Distance Geometry, in Methods in Enzymology, N. J. Oppenheimer and T. L. James, eds., Vol. 177, Academic Press, 1993, pp. 159-204.
19. Moré and Z. Wu, ε-Optimal Solutions to Distance Geometry Problems via Global Continuation, in Global Minimization of Non-Convex Energy Functions: Molecular Conformation and Protein Folding, P. M. Pardalos, D. Shalloway, and G. Xue, eds., American Mathematical Society, 1996a, pp. 151-168.
20. Moré and Z. Wu, Smoothing Techniques for Macromolecular Global Optimization, in Nonlinear Optimization and Applications, G. Di Pillo and F. Gianessi, eds., Plenum Press, 1996b, pp. 297-312.
21. J. Moré and Z. Wu, Global Continuation for Distance Geometry Problems, *SIAM J. Optim.*, Vol. 7, No. 3, 1997a, pp. 814-836.
22. J. Moré and Z. Wu, Issues in Large Scale Global Molecular Optimization, in Large Scale Optimization with Applications, L. Biegler, T. Coleman, A. Conn and F. Santosa, eds., Springer-Verlag, 1997b, pp. 99-122.
23. J. Moré and Z. Wu, Distance Geometry Optimization for Protein Structures, *J. Global Optim.* 15, 1999, pp. 219-234.
24. J. B. Saxe, Embeddability of Weighted Graphs in K-Space Is Strongly NP-Hard, *in Proc. 17th Allerton Conference in Communications, Control and Computing,* 1979, pp. 480-489.
25. Trosset, Applications of Multidimensional Scaling to Molecular Conformation, *Computing Sciences and Statistics* 29, 1998, pp. 148-152.
26. J. Yoon, Y. Gad, and Z. Wu, Mathematical Modeling of Protein Structure with Distance Geometry, to appear in Numerical Linear Algebra and Optimization, Y. Yuan et al, eds, Scientific Press, 2002.

CHAPTER 3. A RIGID GEOMETRIC BUILD-UP ALGORITHM

A paper to be submitted with the complete name (a rigid geometric build-up algorithm for solving the distance geometry problems with sparse distance data)

Di Wu and Zhijun Wu

Abstract

The determination of a protein structure requires solving a so-called distance geometry problem, given a set of distances. With sufficient distance data, the general geometric build-up algorithm can determine a protein structure efficiently, or even in linear time $O(n)$. In this approach, the coordinates of the atoms in a protein are determined one atom at a time, with the distances from four base atoms to the atom to be determined. However, the requirement for four base atoms for the unique determination of each atom is sufficient, but not necessary, and is even redundant for rigid determination. Here we introduce a so-called rigid geometric build-up algorithm, which requires only three instead of four base atoms for the determination of each atom, and can generate rigid and sometimes, even unique structures for very sparse distance data. The algorithm may produce multiple structures, due to the possible reflection for each atom. It keeps track of all combinations and determines a set of structures that are allowed. We present the results obtained by using this algorithm for the determination or generation of the structures for a set of model proteins, and suggest or demonstrate the great potential of using the algorithm for protein structural analysis and determination. In the end, we propose a potential method of protein structure assemble using rigid determination.

Keywords Protein structure determination, distance geometry, geometric build-up

Introduction

Molecular distance geometry problem has important applications in many biological fields including nuclear magnetic resonance (NMR) protein structure determination and protein structure prediction [1]. In general, the distances for certain pairs of atoms in a protein can often be obtained based on our knowledge of various types of bond-lengths and bond-angles [2-3], or through homology modeling and structural alignment [4], or from physical experiments such as nuclear magnetic resonance (NMR) [5]. Therefore, it requires the solution of a mathematical problem called the distance geometry problem to determine the coordinates of the atoms, using given distances. However, the distance geometry problem has been proved to be computationally intractable in general. Especially, in practice, only sparse and inexact distance data is often available, and sometimes distance inconsistencies or errors may also exist. The distance geometry problem can be formalized in different ways and several algorithms have been developed to solve this problem, for example, the SVD and embedding algorithm by Crippen and Havel [6], the CNS partial metrication by Brunger [7], the alternating projection algorithm by Glunt and Hayden [8], the graph reduction algorithm by Hendrickson [9], the multi-scaling algorithm by Trosset [10] and Kearsly, Tapia, and Trosset [11], and the global smoothing algorithm by Moré and Wu [12]. Most of these algorithms can provide an approximate solution to the problem and are very expensive in computation as well.

Recently, a novel approach called the geometric build-up method by Dong and Wu has been developed and can efficiently determine the protein structure with sufficient sparse distances [13-14]. In this approach, the coordinates of the atoms in a protein are determined one atom at a time, using distances from four base atoms to the atom to be determined. For instance, if the coordinates of four atoms which are not in the same plane and the distances from the fifth atom to these atoms are given, then the coordinate of the fifth atom could be determined uniquely by solving simple linear system equations. This simple geometric fact can be easily verified in 3D Euclidean space. Therefore, if all exact distances between all pairs of atoms in a protein are given, the protein structure can be determined with the coordinates of n atoms in n steps, or in order of n floating point operations, but in general, singular value decomposition algorithm requires $O(n^2)\sim O(n^3)$ floating point operations. Note that during the determination, for each atom, the set of four base atoms could be the same or chosen differently corresponding to all exact distance data or sparse exact distance data respectively. However, the application of the general geometric build-up algorithm to sparse distance data is very sensitive to the numerical errors generated in calculation. The reason is clear that the errors in the previously determined atoms are passed to and accumulated in later determined atoms. Especially, for a large system with thousands of atoms, the numerical error can blow up and a protein structure incorrectly determined. To solve the numerical stability problem existing in the sparse distance data,

the algorithm has also been further modified and incorporated the step of rebuilding base atoms, which is called an updated geometric build-up algorithm [15]. This algorithm initializes the coordinates of the base atoms whenever necessary and possible, and can keep the errors from passing over to the atoms to be determined and resulting in incorrect structures. Therefore, the coordinates for many of the atoms can be more accurate, and the errors can be prevented from generating incorrect structures.

Nevertheless, the investigation on sufficient and necessary conditions in solving distance geometry problem remains challenging. In the general geometric build-up algorithm, the requirement for four base atoms for the unique determination of each atom is considered as a strongly sufficient condition, and a protein structure hence could be uniquely determined if this sufficient condition is always satisfied in the determination of each atom. On the other hand, the necessary condition in protein structure determination is found to be that each atom must have at least three distances from other atoms if the protein structure can be determined in 3D Euclidean space. However, the exactly sufficient and necessary condition is still unknown, and the answer to it becomes very important and valuable since it can provide the criteria to justify what distances are necessary or redundant in protein structure determination. Obviously, in the general geometric build-up algorithm, the strongly sufficient condition could be further reduced to the weakly sufficient condition in a sense that the additional atom could be still possibly determined with multiple positions, which is so called rigidity or rigid determination in graph embedding [9]. In that way, a protein structure could be determined rigidly with multiple conformations when given very sparse distance data.

Here, we introduce a so-called rigid geometric build-up algorithm for solving distance geometry problem with sparse exact distance data, and developed the weakly sufficient condition as well as the rigid determination. This algorithm has been shown to be applicable to very sparse distance data while the general geometric build-up method sometimes fails. The idea of this method is based on the fact that using three base atoms is sufficient enough to determine an additional atom, and therefore the original requirement is further reduced to the current weak condition. Each atom can be determined through the simple calculation of a small system of algebraic equations, given coordinates of three base atoms and its distances from them. The implementation of this algorithm has mixed conditions including both four-base-atom and three-base-atom cases, and the unique determination is always preferred. It also allows the determination of an atom if only three base atoms available. Compared to the general geometric build-up method, it requires fewer distances, but an additional atom can be determined rigidly with two possible conformations due to reflection. The algorithm may, therefore, generate a set of structures satisfying given distances for a protein. To control

numerical errors, the idea of the updated geometric build-up method has been implemented as well whenever it is necessary and possible to rebuild three or four base atoms. Another interesting problem is that due to insufficient distance data the protein structure can not be determined sequentially using rigid determination. Using different starting atoms, we can only determine some parts of a protein structure independently. If any two of those parts has more than three common atoms, then we could assemble these two parts, and possibly recover the structure eventually. We describe this idea and specifically study one case.

The general geometric build-up algorithm

The idea of the general geometric build-up algorithm for solving the molecular distance geometry problem is based on the simple geometric relationship between determined and undetermined atoms. This algorithm could determine the protein structure in n steps given the exact distances between all pairs of n atoms in the molecule.

Figure 7. The outline of the general geometric build-up algorithm

The Geometric Build-up Algorithm for Problems with All Exact Distances*

1. Initialize four base atoms that are not in the same plane;
   Find the coordinates of the base atoms with the distances among them.
2. For each of the remaining atoms,
   determine the coordinates of the atom with its distances to the base atoms.
3. All atoms are determined.

*The outline of the general geometric build-up algorithm, given all exact distances (Dong and Wu)

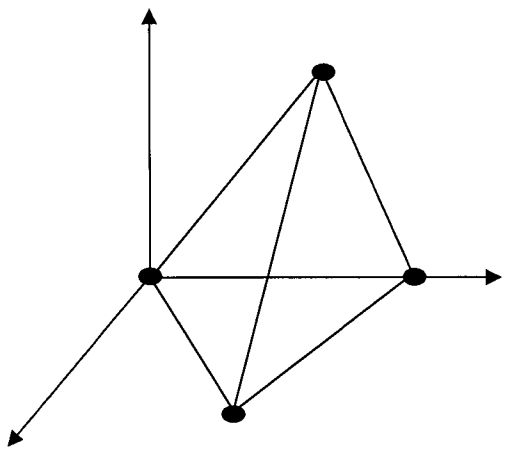
The algorithm has two parts. First, select four initial atoms which are not in the same plane and use the distances among them to find a set of coordinates for them. This is based on the fact that the coordinates of four atoms can be found if all distances among them are available and they are not in the same plane in 3D Euclidean space. Second, use the distances from the atoms to these four atoms to determine the coordinate of each remaining atom, corresponding to that the coordinates of an atom can be fixed if the distances from the atom to four determined atoms are available. These four atoms are so called metric base atoms in the algorithm. In each part, it only requires simple algebraic calculations to determine those coordinates, and especially, for the second part, it needs to

solve for the solution of a small system of linear equations. The algorithm is outlined in figure 7 and a detailed and rigorous description of this method is addressed here.

Theorem 2.1. In the 3D Euclidean space, if the distances among four atoms which are not in the same plane are available, the coordinates of the atoms then be determined with given distances, subject to translation, rotation, and reflection (Wu D and Wu Z 2006) [16].

Proof. See Wu D and Wu Z 2006.

Figure 8. The determination of base atoms



This idea could be also seen from the figure 8. Mathematically, the coordinates of four atoms could be determined through the following steps, given the distances among them:

Let $x_i = (u_i, v_i, w_i)^T$, $i = 1, 2, 3, 4$, be the coordinate vectors of the four atoms. Let $d_{i,j}$ be the given distances between atoms i and j for $i, j = 1, 2, 3, 4$. Without loss of generality, the first atom could be located at the origin of the system, the second on its x-axis, and the third on its xy-plane. Therefore, in this system, we have

$$u_1=0,\ v_1=0,\ w_1=0,$$
$$v_2=0,\ w_2=0,$$
$$w_3=0.$$

We also directly let $u_2=d_{2,1}$ since the distance from the second atom to the first atom is equal to $d_{2,1}$. Then the first two atoms are determined.

Based one the given distances from the third atom to the first and second atoms, respectively, we set up the following equations,

$$u_3^2 + v_3^2 = d_{3,1}^2$$
$$(u_3 - u_2)^2 + v_3^2 = d_{3,2}^2.$$

which could be solved for u_3 and v_3 using following formula,

$$u_3 = (d_{3,1}^2 - d_{3,2}^2)/(2u_2) + u_2/2$$
$$v_3 = \pm(d_{3,1}^2 - u_3^2)^{1/2},$$

and the third atom is then determined completely. In general, either positive or negative of v_3 can be chosen.

Finally, using all the distances from the fourth atom to the first three determined atoms, we can set up three equations,

$$u_4^2 + v_4^2 + w_4^2 = d_{4,1}^2$$
$$(u_4 - u_2)^2 + v_4^2 + w_4^2 = d_{4,2}^2$$
$$(u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2 = d_{4,3}^2.$$

The coordinates u_4, v_4, w_4 for the fourth atom can then be determined by solving the equations, and

$$u_4 = (d_{4,1}^2 - d_{4,2}^2)/(2u_2) + u_2/2$$
$$v_4 = (d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2)/(2v_3) + v_3/2$$
$$w_4 = \pm(d_{4,1}^2 - u_4^2 - v_4^2)^{1/2}.$$

Theorem 2.3. In the 3D Euclidean space, if the coordinates of four atoms not in the same plane and the distances from the fifth atom to the four atoms are given, then the coordinates of the fifth atom can be determined uniquely (Wu D and Wu Z 2006) [16].

Proof. See Wu D and Wu Z 2006.

Figure 9 graphically shows how the method works. The following shows the mathematical calculations.

Let $x_i = (u_i, v_i, w_i)^T$, $i = 1, 2, 3, 4$, be the coordinate vectors of the first determined four atoms and $x_j = (u_j, v_j, w_j)^T$ the coordinate vector of the unknown fifth atom with an arbitrary index j. Let $d_{i,j}$ be the given distances from any of the first four atoms i to the fifth atom j for $i = 1, 2, 3, 4$. We then have the following equations,

$$\| x_1 - x_j \| = d_{1,j}$$
$$\| x_2 - x_j \| = d_{2,j}$$
$$\| x_3 - x_j \| = d_{3,j}$$
$$\| x_4 - x_j \| = d_{4,j}$$

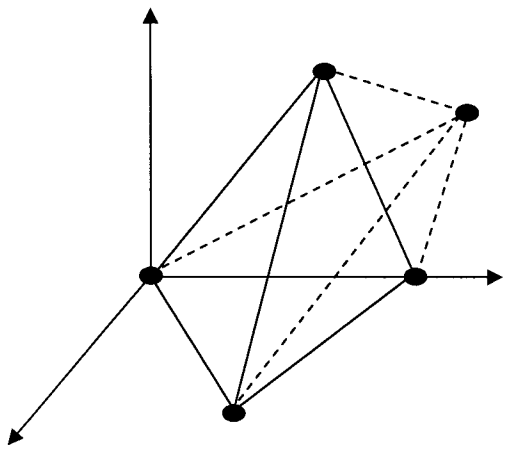
which is equivalent to

$$\| x_1 - x_j \|^2 = \| x_j \|^2 - 2x_j{}^T x_1 + \| x_1 \|^2 = d_{1,j}{}^2$$
$$\| x_2 - x_j \|^2 = \| x_j \|^2 - 2x_j{}^T x_2 + \| x_2 \|^2 = d_{2,j}{}^2$$
$$\| x_3 - x_j \|^2 = \| x_j \|^2 - 2x_j{}^T x_3 + \| x_3 \|^2 = d_{3,j}{}^2$$
$$\| x_4 - x_j \|^2 = \| x_j \|^2 - 2x_j{}^T x_4 + \| x_4 \|^2 = d_{4,j}{}^2$$

For instance, we subtract the first equation from the rest ones and obtain the followings,

$$2x_j{}^T (x_1 - x_2) = (\| x_1 \|^2 - \| x_2 \|^2) - (d_{j,1}{}^2 - d_{j,2}{}^2)$$
$$2x_j{}^T (x_1 - x_3) = (\| x_1 \|^2 - \| x_3 \|^2) - (d_{j,1}{}^2 - d_{j,3}{}^2)$$
$$2x_j{}^T (x_1 - x_4) = (\| x_1 \|^2 - \| x_4 \|^2) - (d_{j,1}{}^2 - d_{j,4}{}^2)$$

Figure 9. The idea of geometric build-up algorithm



In matrix form, the equations are reduced to

$$Ax_j = b_j$$

where

$$A = 2\begin{pmatrix} x_1 - x_2 \\ x_1 - x_3 \\ x_1 - x_4 \end{pmatrix}, 3\times 3$$

and

$$b_j = \begin{pmatrix} (\| x_1 \|^2 - \| x_2 \|^2) - (d_{j,1}{}^2 - d_{j,2}{}^2) \\ (\| x_1 \|^2 - \| x_3 \|^2) - (d_{j,1}{}^2 - d_{j,3}{}^2) \\ (\| x_1 \|^2 - \| x_4 \|^2) - (d_{j,1}{}^2 - d_{j,4}{}^2) \end{pmatrix}$$

If the metric base atoms are not in the same plane, then A will be nonsingular since x_1-x_2, x_1-x_3, x_1-x_4 are linearly independent. Therefore, x_i can be uniquely determined by solving the simple lineal system equations. For a molecule having n atoms and all exact inter-atomic distances, we only need to repeatedly solve the equations for at most n steps. Then the protein structure could be determined using this linear time algorithm given exact all distances.

Figure 10. The outline of general method with sparse exact distances

The Geometric Build-Up Algorithm for Problems with Sparse Exact Distances*

1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. Repeat:
   For each of the remaining atoms,
      find four determined atoms that can serve as its base atoms;
      determine the coordinates of the atom with its distances to the base atoms.
   End
   If no atom is determined in the whole loop, stop.
3. All atoms are determined.

*The outline of the general geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distances (Dong and Wu 2002) [14]

In general, only a subset of all distances among the atoms could be available. Such problem is called distance geometry problem with sparse exact distances. For all exact distance data, the metric base atoms can be unique during the entire determination since the distances from all other remaining atoms to metric base atoms are available. But if only a subset of all distances is provided, the choice of metric base atoms working initially may not be applicable to other atoms later. However, for an atom to be determined, if the previous metric base atoms are not applicable and another group of four determined atoms can be found to serve as new metric base atoms, this atom can still be fixed using the same equations described above. Therefore, the algorithm can be modified to accommodate the sparse case through determining the coordinates of an atom using any determined atoms whenever they can server as its base atoms. Of course, given sparse distance data must be sufficient enough. The modified algorithm is shown in figure 10.

Even though in practice, some sparse distance data problems could be very efficiently solved, the algorithm outlined in figure 2 does not guarantee solving the problem in linear or even polynomial degree time. Note that the for-loop in the second step may be repeated many times until all remaining atoms are determined. A numerical problem in this algorithm for sparse distance data was pointed out [14] and is that the errors introduced in previous numerical calculations are passed to the current atom since the base atoms that are used to determine an atom are determined themselves by some other base atoms in previous steps. This may cause a completely incorrect result. Even though the choice of base atom is based on the determinant of A matrix, it only resolved some of the test cases, but not for all.

An updated geometric build-up algorithm

In this algorithm, two new strategies are implemented to control and minimize the errors introduced in the coordinate calculations, compared to the general geometric build-up algorithm. First, we use condition number instead of the determinant of matrix A in each of the linear systems in the algorithm to decide if the set of base atoms should be chosen or not. For the determination of each atom, we try to find a group of base atoms in all possible combinations so that its condition number is the smallest, corresponding to that an ill-conditioned matrix A with large condition number is numerically instable in solving linear system equations. Second, we always prefer the metric base atoms with all the distances among them, and then re-calculate or re-initialize their coordinates using the methods described in Theorem 2.1. Since they are freshly determined and independent of the coordinates of previously determined atoms, the re-calculated coordinates of them and the coordinates of the additional atom generated using them should have less errors and reduce the error accumulation and delivery. In order to put the atoms back to the original structure by aligning their new coordinates with the old ones, we use regular RMSD calculations to obtain the translation vector and the rotation matrix.

The mathematical description of implementing updated method is following. The new coordinates of metric base atoms could be built through Theorem 2.1. Hence the coordinates of the additional atom could be determined using the new coordinates of metric base atoms through Theorem 2.2. To find the translation vector and rotation matrix, we can do the following steps. Let X and Y be the coordinate matrices of old and new coordinates of metric base atoms respectively. Then, first we compute the geometric centers, xc and yc of the two structures,

$$xc = \frac{1}{4}\sum_{i=1}^{4} X(i,:),\ yc = \frac{1}{4}\sum_{i=1}^{4} Y(i,:).$$

We update matrix Y,

$$Y(1:4,1) = Y(1:4,1) - [yc(1) - xc(1)],$$
$$Y(1:4,2) = Y(1:4,2) - [yc(2) - xc(2)],$$
$$Y(1:4,3) = Y(1:4,3) - [yc(3) - xc(3)].$$

Now the two structures have the same geometric center. We then compute the matrix $C=Y^TX$ and its singular-value decomposition $C=U\Sigma V^T$. Let $Q=UV^T$, and it is easy to verify that $\| X - YQ \|_F / \sqrt{4}$ is then minimized. Therefore Q is the rotation matrix which gives the new coordinates best aligned with the old ones. Then, we can replace X by YQ and determine an additional atom using updated coordinates of base atoms. The outline of this algorithm is shown in figure 11.

Figure 11. The out line of the updated geometric build-up algorithm

The Updated Geometric Build-Up Algorithm for Problems with Sparse Exact Distances*

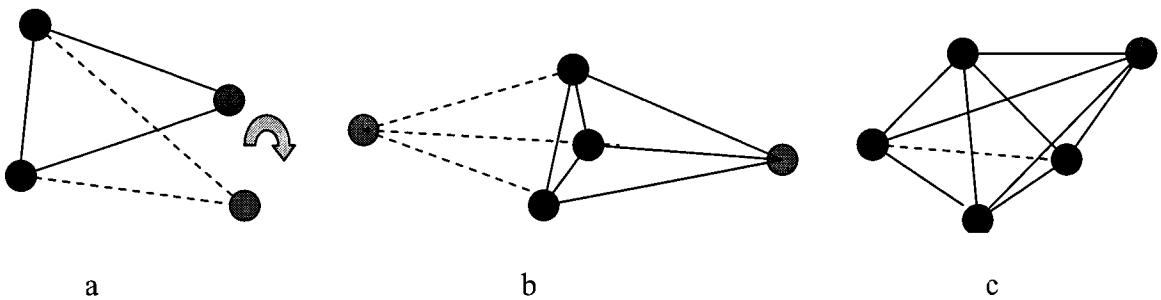
1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. Repeat:
   For each of the remaining atoms,
   find four determined atoms that can serve as its base atoms;
   If four determined atoms are found having all distances among them,
   re-initialize the coordinates of the four atoms;
   put the atoms back to the original structure.
   End
   determine the coordinates of the atom with its distances to the base atoms.
   End
   If no atom is determined in the whole loop, stop.
3. All atoms are determined.

*The outline of the updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distances (Wu D and Wu Z 2006) [16]

A rigid geometric build-up algorithm

Rigidity or Rigid actually comes from graph theory. If an atom has infinite possible positions in its determination, then we call it flexible determination; if it has finite possible (or unique) positions in its determination, then we call it rigid (or unique) determination. For the entire protein with a set of distances, if the protein has finite possible (or unique) conformations in its determination, then the protein is rigid in determination; if it has infinite possible conformations in its determination, then the protein is flexible in determination. In practice, we always care the latter more since rigid determination can still provide the possibly to choose the correct conformation and are more interesting to biologists. Figure 12 shows the detailed explanation.

Figure 12. Flexibility, rigidity and uniqueness*



a b c

*Flexibility VS Rigidity VS Uniqueness. 14(a) shows an example of flexibility that we use coordinates of two atoms (black) and their distances from the third atom (red) to determine the third atom, hence the set of solutions of the coordinates of the third atom can form a sphere. In 14(b), if we increase to three known atoms, then the fourth atom could be determined rigidly, having two possible solutions due to the reflection. In 14(c), this is exactly the idea of the Theorem 2.2, the unique determination.

A rigid geometric build-up method is modified from the general geometric build-up algorithm, incorporating the strategy dealing with the case of rigid determination. In protein structure determination, if the atom has four metric base atoms, then we still use the method in theorem 2.2 to determine the atom uniquely; however, if the atom has distances from only three determined atoms which are not in the same line, then here we would rather determine this atom and obtain two possible sets of coordinates because of the reflection, and keep track of both possible conformations. Three base atoms here are called weak metric base. Ideally, a protein structure determination problem using this algorithm requires fewer distances and it hence could handle even sparser distance data. On the other hand, during protein structure determination, this method may solve substructures rigidly with a

set of possible conformations. Another necessary step is checking the consistence in each possible conformation of substructures right after the determination of atom, using additional distances from this atom to other determined atoms. Only conformations satisfying all available distances between pairs of determined atoms are kept for the next iteration. This algorithm is outlined in figure 5. A detailed explanation will be stated in the following theorem.

Theorem 4.1. In the 3D Euclidean space, if the coordinates of three atoms not in the same plane and the distances from the forth atom to the three atoms are given, then the coordinates of the forth atom can be determined rigidly, for instance, having two possible solutions due to the reflection.

Proof. Let $x_i = (u_i, v_i, w_i)^T$, i = 1, 2, 3, be the coordinate vectors of the three atoms. Let $x_j = (u_j, v_j, w_j)^T$ be the coordinates of an atom to be determined and $d_{j,i}$, i= 1,2,3 be the distances from three atoms. Without loss of generality, we can modify the coordinates of three base atoms for the convenience of calculations. The first atom could be located at the origin of the system, the second on its x-axis, and the third on its xy-plane. Therefore, we could have new coordinates vectors of the three atoms after translation. Note that some coordinates are zeros after modification.

$$u'_1=0,\ v'_1=0,\ w'_1=0,$$
$$u'_2,\ v'_2=0,\ w'_2=0,$$
$$u'_3,\ v'_3,\ w'_3=0.$$

Using all the distances from the fourth atom to the three base atoms, we can set up three equations,

$$\begin{aligned} u_j^2 + v_j^2 + w_j^2 &= d_{j,1}^2 \\ (u_j - u'_2)^2 + v_j^2 + w_j^2 &= d_{j,2}^2 \\ (u_j - u'_3)^2 + (v_j - v'_3)^2 + w_j^2 &= d_{j,3}^2. \end{aligned}$$

The coordinates u_j, v_j, w_j for an atom can then be determined by solving the equations, and

$$\begin{aligned} u_j &= (d_{j,1}^2 - d_{j,2}^2)/(2u'_2) + u'_2/2 \\ v_j &= (d_{j,2}^2 - d_{j,3}^2 - (u_j - u'_2)^2 + (u_j - u'_3)^2)/(2v'_3) + v'_3/2 \\ w_j &= \pm(d_{j,1}^2 - u_j^2 - v_j^2)^{1/2}. \end{aligned}$$

Therefore, the fourth atom has been determined and has two sets of coordinates, which complete the proof.

Figure 13. The outline of the rigid geometric build-up algorithm

The Rigid Geometric Build-Up Algorithm for Problems with Sparse Exact Distances*

1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. Repeat:
   For each remaining atom;
      find four (or three) base atoms (four is always preferred);
         If distances among them are available,
            reinitialize or rebuild the coordinates of base atoms
            put the atoms back to the original structure.
         determine the coordinates of the atom uniquely (or rigidly);
      If additional distances are available
            For each possible structure
               If inconsistency is found
                  Reject it
               Else
                  Keep it for the next iteration
3. All atoms are determined.

*The outline of the rigid geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distances.

Numerical results

The rigid geometric build-up algorithm is implemented in Matlab (Version 7.0) (see the source code in Appendix A). Some matrix-vector calculations used in the algorithm are completed through the Matlab build-in functions, such as linear systems solvers, estimations of condition numbers of matrix and singular-value decompositions. A set of test problems is generated using known protein structures downloaded from the PDB database [16]. In order to test the algorithm in all different kinds of applications, we consider protein structure determination at both atomic level and residual level. In the atomic level, each of the structures is used to obtain two sets of distances, one including all distances ≤ 4 Å and another including all distances ≤ 5 Å. In the residual level, each of the structures is used to obtain three sets of distances and here only distances between Cα atoms are

considered, one including all distances ≤ 7 Å, another including all distances ≤ 7.5 Å, and the other including all distances ≤ 8.5 Å. For each set of distances, we use rigid geometric build-up method to solve the molecular distance geometry problem. As a control, we also run the general updated geometric build-up method for protein structure determination. For the problem of the structure assemble, we selected the protein with PDB id 1CC7 with 72 residues and 567 atoms and set the cut off distance as 7.2 Å. The rigid algorithm fails in determining this protein structure due to insufficient distances at the residual level, but we could rigidly determine two substructures of the protein. One substructure has been determined with four conformations for amino acid residues 1-12 and 31-72, and the other has been determined with 128 conformations for amino acid residues 13-32. Three common Cα atoms existing in both substructures allow us to assemble the structure using RMSD calculations. All results are compared with the experimentally determined structures of corresponding proteins in terms of RMSD. The selections of these cut-off distances are close to the real applications. Usually in atomic level, NMR experiments can provide distances among hydrogen atoms less than 5 Å apart and the length of chemical bonds are even smaller and can be obtained through the knowledge of chemistry, which are even smaller. On the other hand, in residual level, the cut-off distance for Cα atomic contacts is around 7 Å, and the contact map or distances of Cα atoms of a protein could be obtained using homology modeling and other knowledge-based methods [17]. Therefore, the applications of this algorithm could be potentially conducted for those cases. Of course in practice, only lower and upper bounds of the distances could be given. Here, we only consider problems with exact distances. The extension of the algorithm to distance ranges will be investigated later.

Table 2 contains the results of using the rigid geometric build-up algorithm for solving the general test problems in atomic level. They were also compared with the results of using the general updated geometric build-up algorithm. The first column contains the names of the proteins in the PDB database. The second column contains the numbers of atoms in the proteins. The remaining columns list the results of using the rigid and general algorithm for testing problems with 4 Å and 5 Å cutoff distances respectively. For the rigid geometric build-up algorithm, some proteins were determined rigidly, having multiple conformations satisfying given distances, and we listed the smallest RMSD values as well as the number of conformations. For example, 1ABA has two conformations in its determination using rigid geometric build-up algorithm, given a set of distances ≤ 5 Å, and one of them is very close to the original structure of the protein. However, due to the large number of combinations when using rigid geometric build-up method, during the structure determination, the system did blow up for some proteins, including 1ABA and 1BKR both with a set

of distances ≤ 4 Å. Based on the analysis, given distances of these two proteins are actually sufficient enough for determining structures using rigid geometric build-up method, so we would use "fixable" to mean that they can be determined and some strategies are to be developed to solve them in the future. Overall, when considering testing problems, the rigid geometric build-up method can determine all the proteins completely in 1ABA, 1BKR, 1EJG and 1HYP, and some proteins have multiple conformations including 1ABA and 1BKR; but the general update geometric build-up method did not perform very well with same sets of distances, and in testing problems, only 1EJG and 1HYP are completely determined when using sets of distances ≤ 5 Å and all others just failed. For a set of distances ≤ 5 Å of 1BKR on which the general geometric build-up failed, the rigid geometric build-up method can even determine the structure uniquely, though there is a large number of combinations during structure determination steps. This can be also considered as the numerical evidence to that the requirement of four metric base atoms is even redundant for determining a protein structure uniquely.

Table 2. Results of rigid and general algorithms in atomic level

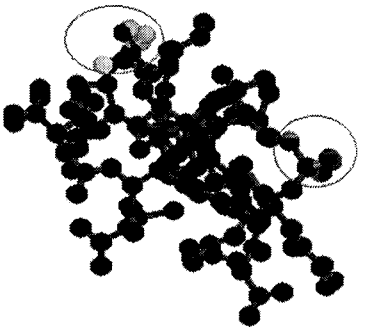
PDB*	Method (Cut off)	4Å RMSD(Å)	5Å RMSD(Å)
1ABA 699	Rigid update	fixable	4.71e-10/2
	Unique update	/	/
1BKR 887	Rigid update	fixable	3.80E-07
	Unique update	/	/
1EJG 637	Rigid update	3.80E-09	9.90E-11
	Unique update	/	8.80E-08
1HYP 656	Rigid update	3.00E-07	1.80E-07
	Unique update	/	2.90E-09

*Results of using the rigid geometric build-up algorithm and the general geometric build-up method to determine protein structures in atomic level. RMSD values are listed and followed by the number of possible conformations if available.

Figure 14 further shows the application of the rigid algorithm to the protein, 1AKG. 1AKG is a small polypeptide containing 16 amino acids and 110 atoms. One thing investigated here is to find the possible minimum cutoff distances for each method including the rigid and the general algorithms, and we start from 5 Å and reduce by 0.5 Å every time until the set of distances less than that cutoff distance is not sufficient. To use the general geometric build-up method to determined 1AKG completely, the possible minimum cutoff distance is 4.5 Å, and the structure could be determined with 8.3×10^{-7} Å in terms of RMSD to the original structure. And the number of distances

used under 4.5 Å is 1638, which is about 14% of all distances. However, using a rigid geometric build-up method to determine protein structure, the cutoff distances can be set as small as 3.5 Å and only about 898, 7.5% of all distances are used, which means 700 distances are removed. Nearly 8192 conformations are found and satisfy the given set of distances and among them the closest conformation to the original structure has RMSD value 4.3×10^{-7} Å. Also from $8192=2^{13}$, it is easy to see that only a few atoms have been determined rigidly due to the reflection, and most of those atoms are in the side chains of some amino acids and found to be located in the surface of the protein with fewer contacts. On the other hand, those atoms especially backbone atoms in the interior of the protein are almost uniquely determined, and hence the determination of the protein is still very descent. Further study of these many possible conformations of 1AKG could incorporate the knowledge of biochemistry and biophysics to identify the native structure, which will be reported later.

Figure 14. The rigid determination of 1AKG*



*1AKG has been determined rigidly, having 8192 possible conformations. Atoms in the circles are those atoms determined rigidly due to the reflection, and different colors represent possible positions. The closest conformation to the original structure has RMSD value 4.3×10^{-7} Å.

The application of the rigid geometric build-up algorithm has also been conducted in protein structure determination in residual level, only considering Cα atoms. Table 3 listed the results of using the rigid and the general geometric build-up algorithms for protein structure determination in residual level. First column contains the PDB names of these proteins downloaded from the PDB database. The second column contains the number of residues in each protein. The last column contains the results of using the rigid and the general algorithms for testing problems with a set of cutoff distances, 7 Å, 7.5 Å and 8 Å respectively. We also listed the number of multiple conformations determined using the rigid algorithm. The rigid algorithm requires fewer distances and

hence can handle some testing problems at smaller cutoff distances. 1EJG and 1IO0 are the only two proteins having been determined using the rigid geometric build-up algorithm with distances $\leq$ 7 Å and 1IO0 have actually 16 possible conformations determined. For a larger cutoff distance 7.5 Å, 1BKR, 1EJG, 1IO0 and 1LIT have been solved using rigid geometric build-up method, and 1IO0 and 1LIT both have two determined conformations respectively. All listed proteins could be determined using the rigid geometric build-up algorithm for a cutoff distance 8 Å, and 1WRI is the only protein determined rigidly having 4 possible conformations. However, the general geometric build-up method can only work for 1BKR, 1EJG, 1IO0 and 1LIT when the cutoff distance is 8Å.

Table 3. The results of rigid and general algorithms in residual level

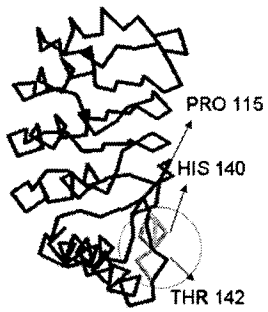
PDB*	Residue	Method (Cut off)	7A	7.5A	8.5A
1BKR	108	Rigid Update	/	2.20e-12	4.30e-12
		Unique Update	/	/	3.60e-11
1EJG	46	Rigid Update	3.60e-14	4.70e-13	1.20e-09
		Unique Update	/	/	7.70e-10
1IO0	166	Rigid Update	6.2e-11/16	3.7e-7/2	6.20e-12
		Unique Update	/	/	6.60e-12
1LIT	131	Rigid Update	/	8.7e-10/2	1.40e-11
		Unique Update	/	/	9.20e-11
1WRI	93	Rigid Update	/	/	5.6e-13/4
		Unique Update	/	/	/

*Results of using the rigid geometric build-up algorithm and the general geometric build-up method to determine protein structures in residual level. RMSD values are listed and followed by the number of possible conformations if available.

Figure 15 illustrates the results of using these two methods to determined 1IO0, which having 166 amino acids. The minimum cutoff distance used in the general geometric build-up method is 8.5 Å, and about 7.5% of all distances, 1886 distances are used. The protein structure has been uniquely determined with high accuracy, 6.0×10^{-12} Å RMSD value deviated from the original structure. On the other hand, the protein structure has been determined rigidly, having 16 possible conformations, when cutoff distance is as small as 7 Å, using the rigid geometric build-up algorithm. It has also shown that using the rigid geometric build-up algorithm, about 500 distances were removed and the percentage of distances used dropped to 5%. Nearly all the backbone Cα atoms are uniquely fixed except Cα atoms of PRO 115, HIS 140 and THR 142 located on the surface of the protein. The Cα atom of PRO 115 has 2 possible conformations, the one of HIS 140 has 4 possible conformations and THR 142 has

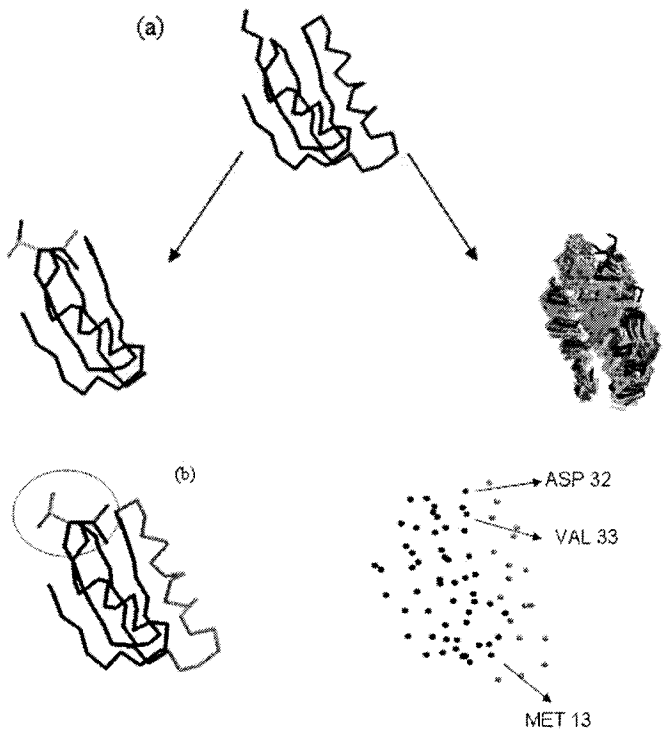
2 possible confirmations, which contribute to 16 combinations including the one closest to the true structure with 6.2×10^{-11} Å.

Figure 15. The rigid determination of 1IO0*



*1IO0 has been determined rigidly, having 16 possible conformations. Nearly all Cα atoms have been uniquely determined expect ones of three residues, including PRO 115, HIS 140 and THR 142. The closest conformation to the original structure has RMSD value 4.3 ×10-7 Å when only comparing the Cα atoms.

Figure 16. Protein structure assembling of 1IO0*



*(a) shows two substructures rigidly determined. The blue one (1-12 and 31-72) has 4 conformations and the red one (13-32) has 128 conformations. (b) shows the structure after alignment with 4 possible conformations

and there are three common Cα atoms of ASP 32, VAL 33 and MET 13 used to accomplish the translation and rotation.

Figure 16 shows the example of protein structure assembling. We select the protein 1CC7 with 72 residues and 567 atoms and set the cut off distance as 7.2 Å. Actually the rigid algorithm fails in determining the protein structure in residual level (only considering Cα atoms), but we could rigidly determine two substructures of the protein. Piece one has been determined with four conformations in the sequence 1-12 and 31-72, and pieces two has been determined with 128 conformations in the sequence 13-32. Three common Cα atoms are ASP32, VAL33 and MET13. Therefore, we could use RMSD calculation to find the translation vector and rotation matrix to align these two structures. Finally about 4 conformations are obtained, in which the closest conformation is 6e-12 Å in terms of RMSD to the true structure.

Conclusions and remarks

A protein structure can be determined by solving a distance geometry problem, given a set of distances. The solution to this problem is not trivial and sometimes even difficult to solve especially when only sparse and distance ranges are given, such as NMR spectroscopy. The molecular distance geometry problem we studied here is considering only sparse but exact distance data. Hence, applying the geometric build-up algorithm to such sparse distance data, the selection of base atoms is dependent on other atoms determined in previous steps in general. However, directly implementing this algorithm caused problems and it was found that the algorithm is numerically instable and sensitive to the numerical errors introduced in calculating the coordinates of the atoms. To control the error accumulation and delivery, the strategy using updated base atoms whenever it is necessary and possible was implemented and could provide more accurate solutions. In the general geometric build-up algorithm, the requirement of four base atoms for the determination of each atom is considered a strongly sufficient condition, and then the protein structure could be uniquely determined. However, such requirement turns out to be redundant not necessary, and for some very sparse distance data, the application of this algorithm has been limited. On the other hand, the investigation of sufficient and necessary condition of solving molecular distance geometry problems becomes very urgent, and the answer to it can be very interesting and valuable to computational scientists as well as experimental biologists.

In this work, we further studied the sufficient condition, incorporated the idea of rigidity and developed a rigid geometric build-up algorithm for dealing the molecular distance geometry with very sparse but exact distance data. We have shown that this algorithm could be applied to sparser distance data while the general algorithm failed. At the same time, multiple conformations for each protein is expected to be determined and satisfying given sparse distances rather than a unique structure in tradition. The key point in this method is reducing the number of base atoms from four to three so that the atom could be still determined with finite positions. Therefore, the rigid geometric build-up algorithm considers both four-base-atom and three-base-atom possibilities, and the unique determination of atoms is always preferred and it also allows the determination of atom if only three base atoms available. The rigid determination of atom results in multiple conformations, and the algorithm will keep track of all possible conformations until the structure is determined. More importantly, in the determination of protein structure, updating base atoms is always performed whenever it is necessary and possible. Compared to the general geometric build-up algorithm, the rigid geometric build-up algorithm can further deal with very sparse distances and determine the protein having multiple conformations satisfying the distance data. However sometimes, proteins could still be determined uniquely using the rigid algorithm while the general fails, which provides the numerical evidence to that the requirement of four metric base atoms for the determination of each atom is redundant even for determining the protein structure uniquely. For a large system with many atoms and very sparse distance data, the number of possible conformations or substructures can be very large and even blow up, which has been seen in our testing problems, hence it still requires further study and additional techniques, especially how to store a huge number of combinations. We also proposed an idea of protein structure assembling using rigid determination. For sparse distance data, if only substructures could be determined independently, there is still a possibility that we could recover the structure based on these common atoms existing in each substructure. The parallel computing technique might be used in the future to advance this study.

To illustrate the idea of geometric build-up as well as the investigations of sufficient and necessary conditions in solving distance geometry problems, we provided a systematic introduction to geometric build-up algorithms, including theorems and important properties. The pseudo codes of all algorithms developed based on the idea of geometric build-up are outlined. The numerical issues and algebraic equations related to the general geometric build-up algorithms are discussed, and updating coordinates of base atoms and the rigid determination as well as checking consistence using additional distances are presented as well. Some numerical results of testing both the rigid and general geometric build-up algorithms are shown, given a set of problems generated with known

protein structures. Particularly, we considered the application to protein structure modeling in both atomic and residual level (considering only C-alpha atoms), and for each level, we selected a set of cut off distances to generate sparse distance data problems. The testing results showed that the rigid geometric build-up algorithm with updating base atoms can further deal with very sparse distance problems and it could determine protein structures rigidly or sometimes even uniquely while the general geometric build-up algorithm failed. For protein structures determined rigidly, having many possible conformations, we expect to incorporate other techniques or knowledge, such as biochemistry, biophysics and potential energy function to further study. An interesting example of protein structure assembling also provides the idea of further application of rigid determination and the possibility of implementing parallel computing.

Actually in practice, the given distances usually are in ranges and very sparse, such as NMR spectroscopy, and the algorithm introduced in this paper may be valuable only theoretically and not be applicable to that case. On the other hand, the sufficient and necessary condition in solving distance geometry problems is further investigated, and the new algorithm incorporating the idea of the rigid determination does show a great potential in applications. The possible extension of this algorithm to sparse distance constraints with lower and upper bounds will be studied later.

Acknowledgements

We would like to thank Dr Robert Jernigan for reading the paper and offering helpful suggestions. The support for the first author from the ISU Graduate Program on Bioinformatics and Computational Biology and ISU department of mathematics is also gratefully acknowledged.

References

1. J. Yoon, Y. Gad, and Z. Wu, Mathematical Modeling of Protein Structure with Distance Geometry, Numerical Linear Algebra and Optimization, Y. Yuan et al, eds, Scientific Press, 2002.
2. C. L. Brooks III, M. Karplus, and B. M. Pettitt, Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, John Wiley & Sons, 1988.
3. T. E. Creighton, Proteins: Structures and Molecular Properties, 2nd Edition, Freeman and Company, 1993.

4. Havel TF, Snow ME. 1991. A new method for building protein conformations from sequence alignments with homologues of known structure. *J Mol Biol* 217:1-7.
5. A. T. Brüger and M. Niles, Computational Challenges for Macromolecular Modeling, in *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publishers, 1993, Vol. 5, pp. 299-335.
6. G. M. Crippen and T. F. Havel, Distance Geometry and Molecular Conformation, John Wiley & Sons, 1988.
7. AT. Brunger, PD Adams, GM Clore, WL DeLano, et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr.* 1998 Sep 1;54 (Pt 5):905-21.
8. W. Glunt and T. L. Hayden and M. Raydan, Molecular Conformations from Distance Matrices, *J. Comput. Chem.*, Vol. 14, No. 1, pp. 114-120, 1993.
9. B. A. Hendrickson, The Molecular Problem: Determining Conformation from Pairwise Distances, Ph.D. thesis, Cornell University, 1991.
10. A. Kearsly, R. Tapia, and M. Trosset, Solution of the Metric STRESS and SSTRESS Problems in Multidimensional Scaling by Newton's Method, *Computational Statistics* 13, 1998, pp. 369-396.
11. M.W. Trosset. Applications of multidimensional scaling to molecular conformation. *Computing Science and Statistics*, 29(1):148-152, 1998.
12. J. Moré and Z. Wu, Smoothing Techniques for Macromolecular Global Optimization, in Nonlinear Optimization and Applications, G. Di Pillo and F. Gianessi, eds., Plenum Press, 1996b, pp. 297-312.
13. Q. Dong and Z. Wu, A Linear-Time Algorithm for Solving the Molecular Distance Geometry Problem with Exact Inter-Atomic Distances, *J. Global Optim.*, Vol. 22, 2002, pp. 365-375.
14. Q. Dong and Z. Wu, A Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse Distance Data, *J. Global. Optim.*, Vol. 26, 2003, pp. 321-333.
15. D. Wu, and Z. Wu. An Updated Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse Distance Data. *J. Global. Optim.,* 2006 (accepted).
16. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, 28 pp. 235-242 (2000)
17. M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2:295--306, 1997.

CHAPTER 4. PIDD: DATABASE FOR PROTEIN INTER-ATOMIC DISTANCE DISTRIBUTIONS

The paper submitted to the Nucleic Acid Research

Di Wu, Feng Cui, Robert Jernigan and Zhijun Wu

Abstract

PIDD is a dedicated database and structural bioinformatics system for distance based protein modeling. The database is developed to host and analyze the statistical data for protein inter-atomic distances based on their distributions in databases of known protein structures such as in the PDB Data Bank. PIDD is capable of generating, caching, and displaying the statistical distributions of the distances of various types and ranges. The collected information can be used to extract geometric restraints or mean-force potentials for protein structure determination including NMR structure determination and comparative model refinement. PIDD is supported with a friendly designed web interface so that users can easily specify the distance types and ranges, and retrieve, visualize, or download the distributions of the distances as they desire. PIDD is freely accessible at http://www.public.iastate.edu/~diwu/pidd.

Introduction

The knowledge on inter-atomic distances in proteins is a valuable source of information for protein structural analysis and structure determination. Protein inter-atomic distances may be detected by using physical experiments such as NMR (nuclear magnetic resonance spectroscopy) (1), or estimated with chemistry knowledge concerning various types of bond lengths and bond angles (2-3). However, in either case, only a small subset of all distances can be obtained for various technical reasons (1, 4). They can only be estimated approximately in certain ranges instead of exact values as well because of inevitable estimation errors. Therefore, obtaining additional distance information beyond the current theoretical and experimental limitations is important yet challenging for the further development of distance based protein modeling.

In this paper, we introduce a computational approach for deriving distance data for proteins based on the distributions of the distances in the databases of known protein structures. In particular, we describe the development of a protein distance distribution database PIDD for calculating and storing the distributions of the distances in databases of known protein structures and using the distribution data to derive distance constraints and mean-force potentials (5) for structural analysis and modeling.

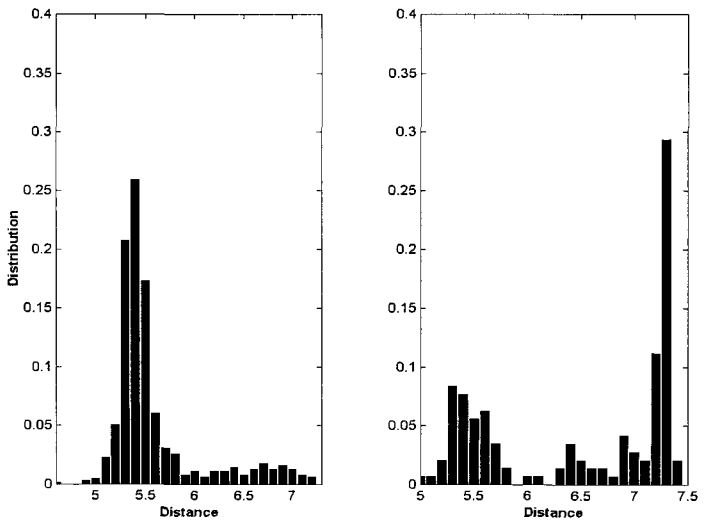
The basic idea of our approach is that in order to estimate the distances for various pairs of atoms, we find all the information for how the distances for different pairs of atoms are distributed in known protein structures. Then, for each distance, we assign a probability according to the distribution of the distances of the same kind. Such probability information can be very useful for evaluating estimated distances or building proper protein conformations. For example, in order to see if 5 Å is a proper distance between C_α in Alanine and C_β in Tryptophan when the two residues are separated by a Cystine, we calculate all the distances of the same type in the known proteins in structural databases and then group the distances according to their lengths. We can then obtain the distribution of this type of distances within a given distance range, say between 0 and 50 Å, where the probability for the distance to be 5 Å can be easily identified. Figure 17 shows more examples of protein inter-atomic distance distributions calculated from databases of known protein structures.

Indeed, based on our calculations on the distributions of the distances in the structures in PDB Data Bank (6), we have found that 1) The majority of short to medium ranged distances are non-uniformly distributed, indicating that proteins do have preferences when forming these distances; 2) as more and more protein structures are determined, good estimations on the distributions of the distances are possible, and they can be obtained with reasonable statistical significances; 3) many distances in low-resolution structures deviate from their average distributions by more than two standard deviations, and in most cases, the deviations are found in under-determined regions of proteins; 4) it follows that distance constraints or mean-force potentials can be derived from the distributions of the distances and be applied to "correct" or "refine" low-resolution structures (7-8).

While the importance of the distance distribution data is easy to justify, the calculation of the data can be daunting, requiring a complete search for the distances in structural databases for each different distance type, while there can be millions of different distance types, defined in terms of the types of the two atoms related to the distance, the types of the two corresponding residues, and the types of the residues separating them in the sequence. Even just storing and managing such an enormous amount of data can be quite challenging. For this reason, we have developed a database system for automatically generating, storing, and analyzing all the distribution data for protein inter-

atomic distances. The system consists of two coupled databases, one called the structural database for storing high-resolution structures downloaded from structural databases, and another called the distance database for storing the distribution data for the distances. The data in the distance database is calculated and collected from the structural database. The distance database can be used by the users to store, query, and analyze the distributions of any distances of interest. At the beginning, only the data for commonly used distance types are computed and stored, to avoid unnecessary space use. If the distributions for certain distances are requested, but have not been pre-calculated and -stored, they are computed from the structural database and stored in the distance database afterwards. In this way, the database can eventually be developed to contain necessary distance distributions, yet does not have to keep all the overwhelming information. The database system is developed using MySQL. Currently, it has 2090 high-resolution structures downloaded from PDB Data Bank and up to 320,000,000 distance distribution records. The system was supported with a friendly web interface so that users can easily specify the distance types and ranges, and retrieve, visualize, or download the distributions of the distances as they desire. It is freely accessible at http://www.public.iastate.edu/~diwu/pidd.

Figure 17. Samples of distance distributions*



*The graph on the left is the distribution of the distances between C_α in TYR and C_α in TYR separated by LYS in sequence. The graph on the right is the distribution of the distances between C_α in SER and C_α in TRP separated by GLY in sequence.

Systems and methods

Data source

When downloading the known protein structures from the PDB Data Bank, we have considered only those containing the chains of amino acids rather than protein complexes such as protein-DNA, protein-RNA, and protein-protein complexes. To obtain more accurate and reliable results, we only downloaded structures determined by X-ray crystallography with resolution higher than 2.0 Å. In future, we will consider including NMR structures as well. To reduce the redundancy in homologous structures, only proteins with sequence similarities less than 70% were used. Based on these criteria, total 2090 qualified protein structures were selected from the PDB Data Bank as of April 12, 2005.

Data structure

PIDD has two levels of databases, one called the structural database and another called the distance database. Both databases are implemented using MySQL. The structural database stores the sequence and structure information for a large set of high-resolution protein structures, with a similar data structure as the structural data represented in the PDB Data Bank. Each record in the structural database is similar to an atom record in the PDB file, but contains a smaller number of fields. It has the PDB name of the protein, the residue name, the index for the atom, the atom name, and the x, y, z coordinates of the atom (see Figure 18). All the PDB files of the downloaded protein structures are converted into this format and stored in the structural database as MySQL database files. By using the MySQL database management system, the structure files can be processed much more efficiently and directly. No special scripts are required to parse the regular PDB text files. The distance database stores the distributions of the distances in known proteins calculated for every different type of distances. The calculations are based on the distributions of the distances in the downloaded structures in the structural database.

In order to obtain the distribution data for the distances of various types and ranges, we specify the distances by using the types of the atoms it involves, the types of the residues containing the atoms, and the types of the residues in between the two terminal residues in sequence. After calculating and collecting all the distances of each distance type from the structural database of PIDD, the statistical distribution of each distance type can be obtained. Let D be the distance between two

atoms, A_1 and A_2. Let R_1 and R_2 be the two residues where A_1 and A_2 are located, respectively. Let S_1, ..., S_N be the residue sequence in between R_1 and R_2. Then, the distribution of the distance D between atoms A_1 in R_1 and A_2 in R_2 where R_1 and R_2 are separated by S_1, ..., S_N can be represented by a distribution function $P[A_1,A_2,R_1,R_2,S_1, \ldots, S_N](D)$ and defined for any D in $[D_i, D_{i+1}]$, where $D_i = 0.1 \times i$ Å, $i = 0, 1, \ldots, n$, to be the number of collected distances of this particular type in $[D_i, D_{i+1}]$, normalized by the total number of collected distances of the same type in all $[D_i, D_{i+1}]$, $i = 0, 1, \ldots, n$.

$$P[R_1,R_2,A_1,A_2,S_1,\ldots,S_N](D) = \frac{\text{Number of distances of this type in } [D_i, D_{i+1}] \ni D}{\text{Number of distances of this type in } [D_0, D_n]}$$

Each record in the distance database therefore contains the distribution data for a particular type of distances, and it has the types of atoms, A_1 and A_2, the types of ending residues, R_1 and R_2, and the types of separating residues, S_1, ..., S_N , that define the type of the distances followed by the number of distances of this type found in each of the distance intervals $[D_i, D_{i+1}]$, $i = 0, 1, \ldots, n$-1.

Figure 18. Data structures of the databases

Structural Database

PDB ID	Residue	Index	Atom	X	Y	Z

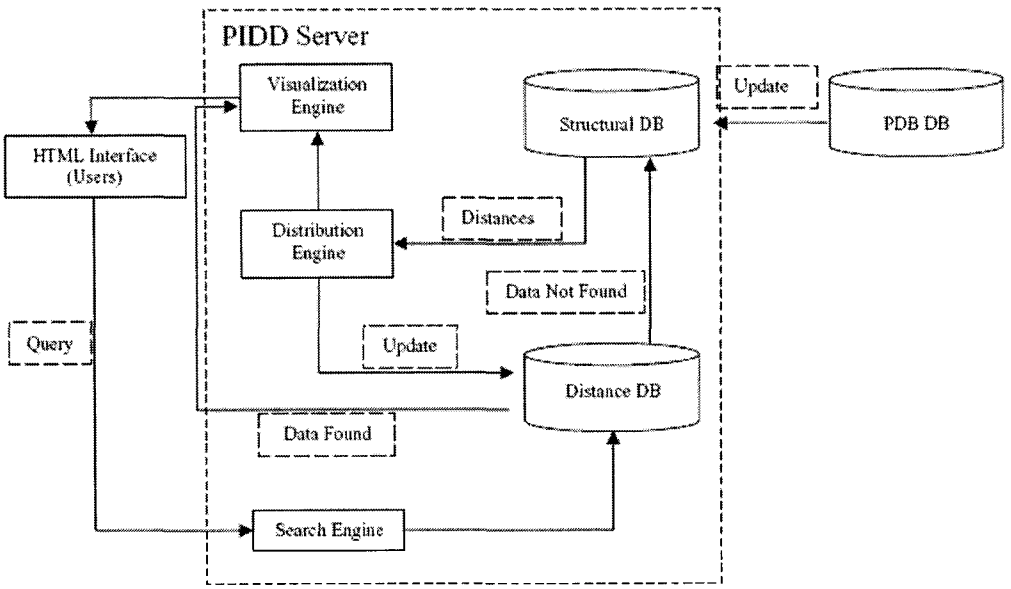
Distance Database

R_1	R_2	A_1	A_2	S_1	...	S_N	$\#D_0$	...	$\#D_{n-1}$

*The above: the record of the atom in the structural database: PDB ID – ID of protein in PDB Databank; Residue – the name of the residue containing the atom; Index – the index for the atom; Atom – the name of the atom; X, Y, Z – x, y, z coordinates of the atom. The bottom: The record for the distribution of the distance, one for each different type: R_1, R_2 – the two residues; A_1, A_2 – the two atoms; S_1, ..., S_N – separating residues; $\#D_i$ – the number of distances in $[D_i, D_{i+1}]$, $i = 0, \ldots, n-1$.

System architecture

Figure 19. The system architecture*



*This automated system generates and processes the data dynamically. The system is implemented in MySQL and Perl. Users could access freely the database at http://www.math.iastate.edu/pidd. It requires specifying and inputting the distance type and then the user could choose to view the graph of distribution function as well as download the related results.

PIDD is implemented with MySQL (see source code in Appendix B). It consists of two databases, a structural database and a distance database, and three computational engines, a search engine, a distribution engine, and a visualization engine. In addition, there is a program written in Perl for automatically downloading the structures from PDB Data Bank and updating the structural database, and a web interface written in HTML for users to gain online access to the system.

The structural database stores the sequence and structure information for a set of high-resolution protein structures. The distance database stores the distribution data for the distances, with one record for one distance type. Since the distance type is defined in terms of the atom types, residues types, and the separating residues, there can be a huge number of distance types, and the amount of distribution data can be enormous. For example, if we assume that there are 10 different atoms types for A_1 and A_2, 20 different residue types for R_1, R_2, S_1, ..., S_N, then even just for the distances with three separating residues ($N = 3$), there are already 320 million possible distance types.

For this reason, we purposely designed the system to have both structural and distance databases so that the distance database can be built dynamically from the structural database. More specifically, at the beginning, we only compute and store the distribution data for some commonly used distance types, which can be queried or processed directly in the distance database. However, if distributions for certain distances that have not been pre-calculated and -stored are requested, they will be computed on fly from the structural database and stored into the distance database afterwards. In this way, the database can eventually be developed to contain all necessary distance distributions, yet does not have to be overwhelmed by the possible combinatorial growth of data, saving both storage space and search time.

The computational engines work together as follows. The search engine takes the query from a user and searches for the distribution of the specified type of distances in the distance database. If the requested distribution has been pre-calculated and -stored in the distance database, the search engine returns it directly. Otherwise, the distances of the specified type will be computed and collected from the structural database and passed to the distribution engine. Based on the collected distances, the distribution engine calculates the distributions of the distances over discrete distance intervals, and saves them in the distance database. The visualization engine is responsible for displaying the requested distribution function through a graphics interface. Figure 19 shows the architecture of PIDD graphically. Note that the structural database can be updated whenever new proteins are deposited into the PDB Data Bank, and the access to PIDD can be done conveniently through a well-designed web interface.

Features

A web user interface is designed so users can gain access to PIDD anywhere online. The interface provides various visualization tools and functions for researchers to display and analyze requested data. The users can obtain helps from the tutorial, references, or related publications available at the website. The tutorial is clearly written and provides many examples (see Appendix C).

The front page of the interface as shown in Figure 20 describes the PIDD system, its design purpose, and user guidelines. More in-depth description about research on database-derived distance constraints and mean-force potentials and distance-based protein modeling is provided on the research page. Links to tutorial, references, and publications are also provided. Currently, the PIDD front page can be reached at http://www.math.iastate.edu/pidd/.

Figure 20. The PIDD frontpage

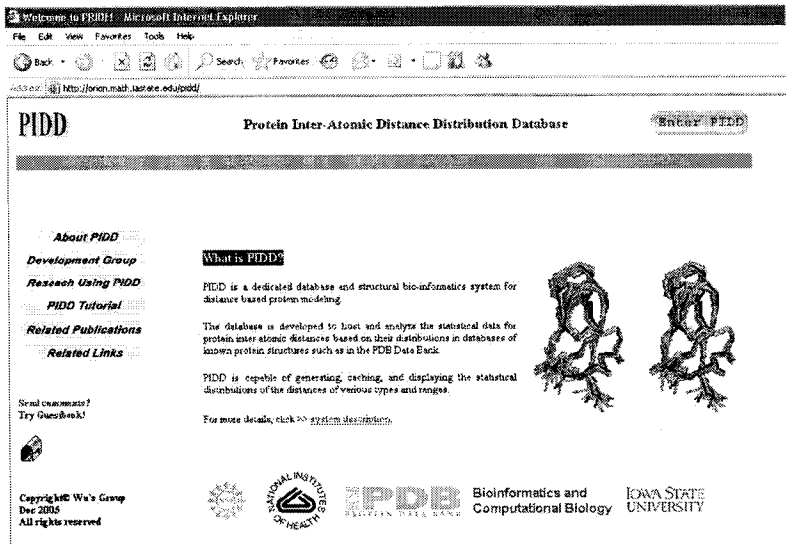
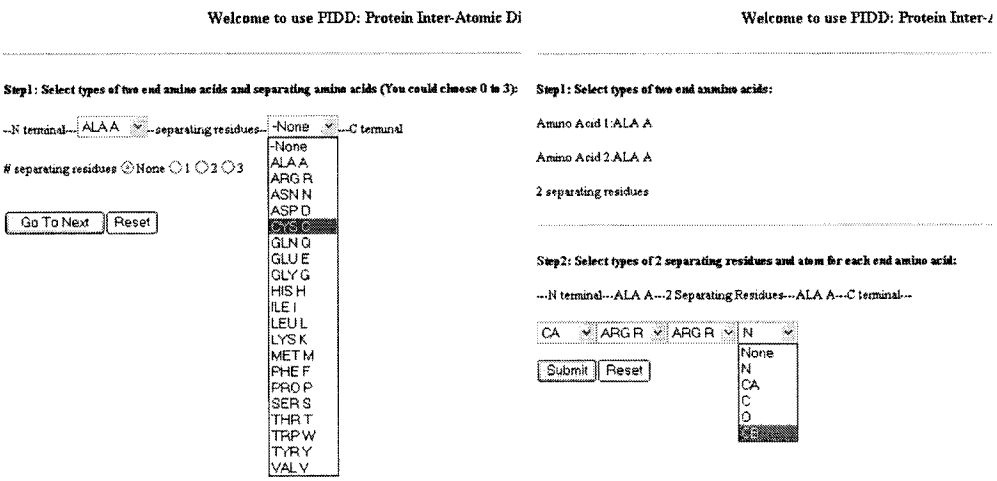
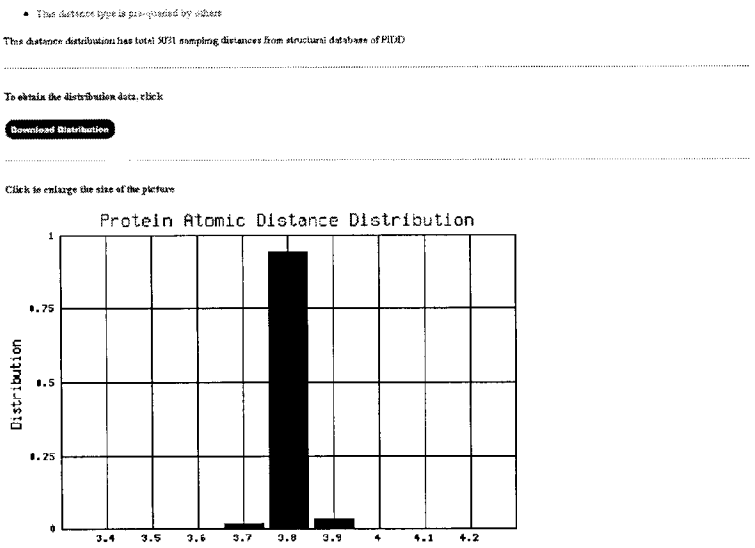


Figure 21. PIDD input selctions



Two web pages are directly related to the use of the PIDD database. One shown Figure 21 allows the users to choose the distance type to be searched for via simple menu selections. Typically, the users follow three selection steps: (i) specify the two end residues and the number of separating residues; (ii) specify the types of the two atoms in the two end residues, and the types of all separating residues; (iii) submit the query. The system returns the distribution of the specified type of distances and displays it in a graph as shown in Figure 22. The current version of PIDD allows the users to specify up to 3 separating residues and handles one distance type per query. It can be used simultaneously by multiple users.

Figure 22. Graphics display



Sample applications

The motivation for the development of PIDD is to provide an easy access to the information on how the inter-atomic distances are formed as revealed in their distributions in known proteins. Such information can be valuable for protein structural analysis, classification, as well as modeling building. In particular, it can be used to extract geometric restraints or mean-force potentials for protein structure determination including NMR structure determination and comparative model refinement.

The distance distribution data has been used to analyze NMR determined structures as reported (9). The inter-atomic distances for 462 averaged and energy-minimized NMR structures downloaded from the Protein Data Bank were examined and compared with their distribution functions (more specifically, for distances between atoms in two residues separated by zero or one residue). The results showed that many of these distances have deviations larger than two standard deviations. For example, the distribution of the distance between C_β in ALA and the carbonyl C in ASP separated by one residue was found to have a mean around 7.1 Å and standard deviation equal to 1.05 Å, while the distance between such pair of atoms across the 20th and 22nd residues in the NMR structure 2GB1 was 4.6 Å, which was 0.37 Å smaller than the mean minus two standard deviations. More example cases of distance deviations in 2GB1 are given in Table 4. In fact, in each of 462 NMR structures, similar deviations were found in 2% to 44%, or in an average of 21.98% of the residue

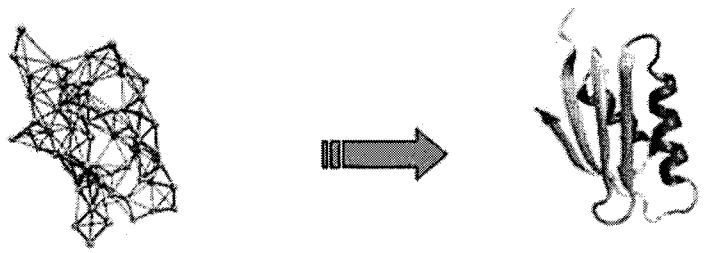
pairs that are separated by one or zero residue along the protein backbone. The deviations were not only found among backbone atoms (N, O, C, C_α), but also between backbone (N, O, C, C_α) and side-chain atoms (C_β). In most cases, the residues having such distance deviations were located on exposed parts of the proteins, which was consistent with the fact that the surface residues are usually of high mobility and more difficult to determine by NMR.

Table 4. Distance deviations in NMR determined structures

Res. No.	Res.1	Atom 1*	Res. No.	Res. 2	Atom 2	Mean	2×STD	Distance
19	GLU	C	20	ALA	C	3.1	0.4	3.62
20	ALA	CB	22	ASP	C	7.1	2.1	4.63
20	ALA	CB	22	ASP	O	7.8	2.5	3.53
21	VAL	N	22	ASP	O	5.9	1.0	4.28
21	VAL	CB	23	ALA	N	5.7	0.9	6.95
22	ASP	CB	23	ALA	C	5.4	0.6	4.69

*Atomic pairs (Atom 1 and Atom 2) across some of the residues (Res .1 and Res .2) in 2GB1 with distances deviating more than two times of their standard deviations (STD) from their average distributions (Mean) in known protein structures.

Figure 23. Distance geometry problems*



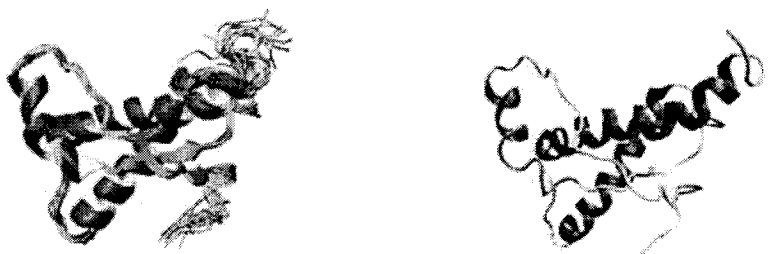
*Given a set of inter-atomic distances, the atomic coordinates and hence the 3D structure of a protein can be determined by solving a so-called distance geometry problem.

An important application of PIDD is structure determination or refinement. A set of distance constraints or mean-force potentials can be obtained by using the distribution data and applied to structure determination and refinement, for example, for NMR determined structures. In general, a set of inter-proton distances of a protein can be obtained by using NMR spectroscopy. The protein structure can then be determined by solving a so-called distance geometry problem (10) (see Figure 23). However, regions in NMR determined structures are often under-determined due to incomplete

or inaccurate distances data (see Figure 24). Overall, the quality and resolution of NMR determined structures are still not as high as those of X-ray crystallographic structures (11).

In order to increase the accuracy of the NMR determined structures, F. Cui et al. (7, 9) and D. Wu et al. (8) used the distributions of the inter-atomic distances in known proteins as calculated in PIDD and derived a set of range constraints for the distances, and applied them to refining a set of NMR determined structures, along with original NMR experimental constraints. The results showed that with addition distance constraints or mean-force potentials, the structures were improved significantly in terms of standard measures, including the energies of the final structures, the Ramachandran plots (12), the RMSD values of the structures compared with X-ray reference structures, etc. For example, for the prion protein E200K the percentage in the most favorable region of the Ramachandran plot was increased from 85% (left) to 90% (right) after the protein was refined by using the database derived distance constraints (9) (see Figure 25).

Figure 24. NMR ensembles of pig prion protein

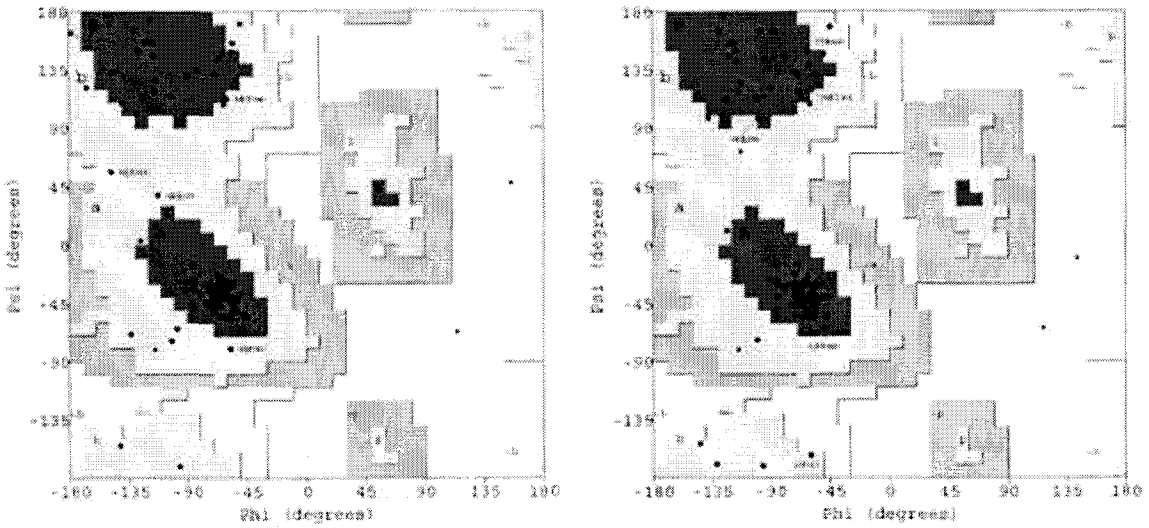


Ensemble of Structures Representative Structure

*NMR determined structures of pig prion protein (residues: 121-231). The left is the ensemble of the accepted structures. The right is the representative structure usually chosen from an average and energy-minimized structure.

It is well-known that NMR determined structures are not as detailed as X-ray crystal structures. The discrepancies between the NMR and X-ray structures may be due to the flexibilities of the NMR structures in solution, while some of them may indeed be caused by the incorrectly formed regions in the NMR models. As indicated in the above applications, the distance distributions generated from PIDD can clearly be used to either find possible errors existing in NMR determined structures or generate additional distance constraints or potentials to refine the structures. There is also a great potential of using the same type of data for refining comparative models.

Figure 25. Ramachandran plots for original and refined E200K*



*After employing additional distance constraints, the Ramachandran plot of NMR determined structures for prion E200K is improved significantly, with 85% of the residues in the most favorable region (left) increased to 90% of the residues in the most favorable region (right).

Future developments

The current version of PIDD provides the basic functions for converting and processing data for protein distance distributions. More tools will be developed to facilitate various tasks of structural analysis including tools for computing the distributions of the distances in more specific structural environment, such as distributions for certain types of distances in alpha helices versus beta sheets. Currently, we have only considered relatively short-range distances with a maximum of three separating residues in primary sequence. In the future, we will also include all statistically significant long-range distance distributions. The reason that we have not considered the distances of all ranges is that many long-range distances either do not have clear distribution patterns or are difficult to sample and analyze. With the increasing number of high-resolution structures being determined, many structural properties, such as torsion angles, inter-atomic distances, residue volumes, side-chain orientations, can all be analyzed from their statistical distributions in known proteins. Therefore, in future, we will extend our work on PIDD to the development of a general protein geometry database that includes the statistical distribution data for many other protein

geometric properties besides the distances. Such a system will be able to provide more complete information on protein conformations and have even greater potentials as bioinformatics tools for protein structural analysis and structural modeling.

Acknowledgements

The authors would like to thank Peter Vedell, Ajith Gunaratne, and Andrew Severin for helpful discussions and valuable suggestions. The work is partially supported by the research funds provided by the Department of Mathematics, the Graduate Program on Bioinformatics and Computational Biology, and the Lawrence Baker Center for Bioinformatics and Biological Statistics at Iowa State University.

References

1. Wuthrich,K., NMR of Proteins and Nucleic Acids, Wiley, New York, 1986.
2. Brooks, CL., Karplus, M., Pettitt, M., Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, John Wiley & Sons, 1988.
3. Creighton, TE., Proteins: Structures and Molecular Properties, 2nd Edition, Freeman and Company, 1993.
4. Clore, GM., Gronenborn, AM., Determination of three-dimensional structures of proteins in solution by nuclear magnetic resonance spectroscopy. *Protein Eng.* 1987 1(4):275-88.
5. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, MJ.. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J Mol Biol.* 1990, 5;216(1):167-80.
6. Bernstein, FC., Koetzle, TF., Williams, GJ., Meyer, EF Jr., Brice, MD., Rodgers, JR., Kennard, O., Shimanouchi, T., Tasumi, M., The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol.* 1977, 25;112(3):535-42
7. Cui, F., Mukhopadhyay, K., Young, WB., Jernigan, R., Wu, Zj., Refinement of Underdetermined Loop Regions of Prion Proteins by Database Derived Distance Constraints, submitted, 2006.
8. Wu, D., Jernigan, R., Wu, Zj., Refinement of NMR-Determined Protein Structures with Database Derived Mean Force Potentials, in preparation, 2006.

9. Cui, F., Jernigan, R., Wu, Zj., Refinement of NMR-determined protein structures with database derived distance constraints. *J Bioinformatics and Computational Biology,* 2005, 3(6):1315-29.
10. Crippen, GM., Havel, TF., Distance Geometry and Molecular Conformation, Research Studies Press, UK 1988
11. Doreleijers, JF., Rullmann, JA., Kaptein, R.. Quality assessment of NMR structures: a statistical survey. *J Mol Biol.* 1998, 7;281(1):149-64.
12. Ramachandran GN, Sasiskharan V. Conformation of polypeptides and proteins. *Adv. Protein Chem.,* 1968, 23:283-437.

CHAPTER 5. REFINEMENT OF NMR-DETERMINED PROTEIN STRUCTURES WITH DATABASE DERIVED POTENTIALS

The paper to be submitted

Di Wu, Robert Jernigan, and Zhijun Wu

Abstract

Due to the limited distance data available from the experiments, the structures determined by NMR Spectroscopy may not always be as accurate as desired. Further refinement of the structures is often required and sometimes critical. With the increase of high quality protein structures determined and deposited in PDB Data Bank, commonly shared protein conformational properties can be extracted based the statistical distributions of the properties in the structural database and used to improve the outcomes of the NMR-determined structures. Here we examine the distributions of protein inter-atomic distances in known protein structures. We show that based on these distributions, a set of mean-force potentials can be defined for proteins and employed to refine the NMR-determined structures. We report the test results on 70 NMR-determined structures and compare the potential energy, the Ramachandran plot, and the ensemble RMSD of the structures refined with and without using the derived mean-force potentials.

Keywords NMR protein structure refinement; protein structure database; Ramachandran plot; mean force potentials; structural bioinformatics

Introduction

NMR (Nuclear Magnetic Resonance Spectroscopy) is a major experimental technique available for protein structure determination [1]. There are about 30,000 structures determined and deposited in PDB Data Bank now. About 15% of them are determined by NMR [2]. The advantage of using NMR is that the protein does not need to be crystallized (which can be difficult and time

consuming) and the structure can be determined in solution (a factor sometimes indispensable for proper folding). NMR can also be used to obtain certain dynamic properties of proteins such as the flexibilities of the proteins in solution. However, similar to other techniques, due to the limited data that can be obtained from the experiment, the structures determined by NMR are not necessarily always as accurate as desired and further refinement of the structures is often required [3].

The most common types of conformational constraints that can be obtained from NMR include the distances between hydrogen atoms estimated via Nuclear Overhauser Effects (NOE) and the dihedral angles around certain bonds through J-coupling [4]. The NOE intensity for two magnetically interacting hydrogen atoms is inversely proportional to the sixth power of the distance between the atoms and can therefore be detected only for atoms in very short distance (< 5Å). In other words, only the distances less than 5Å between hydrogen atoms may be estimated through NOE. Also, the NOE intensity cannot be detected so accurately. Usually it is reduced by other interactions and becomes weaker when detected, and therefore, only a rough upper bound for the distance may be obtained. The lower bound for the distance can be determined for example by using the Van der Waals radii of the atoms. With these distance constraints (along with the estimations on some of the dihedral angles around flexible bonds), an ensemble of structures (other than a single structure) whose distances are within the estimated ranges can then be determined (Figure 26). While the structures determined by NMR are not as exclusive as other types of structural models, the variation of the structures in the ensemble somehow correlates the flexibility of the structures in solution and can often be used to show the dynamic behavior of the protein [5-6].

Figure 26. NMR determined structures of pig prion protein*

* Typical NMR determined structures. Shown are the ensemble of structures (left) and the averaged and energy-minimized structure (right) of the pig prion protein determined by NMR.

To obtain a meaningful ensemble of structures, it is important to have a sufficient set of distance constraints. Depending on how many and how accurate the distance bounds are available, the quality of the structures often varies [7-8]. Further refinement of the structures is always required,

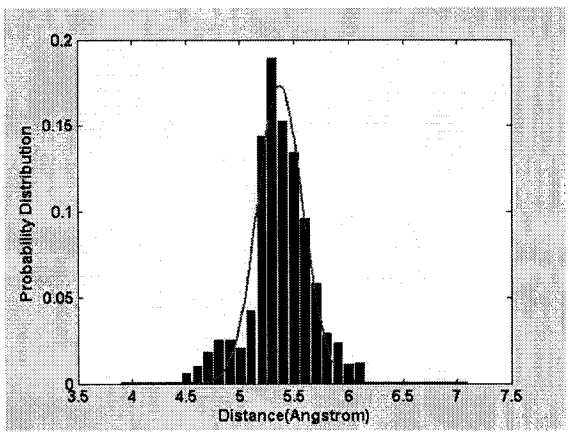
with either theoretical approaches (e.g., energy minimization [9]) or experimental options (e.g., dipolar coupling [10-11]). Knowledge-based approaches have also been employed, including deriving additional dihedral angles or distance constraints from known protein structures [12-15]. Indeed, with rapid increase in both number and quality of protein structures being determined and deposited into PDB Data Bank [2], many structural properties such as secondary structure motifs, native contact patterns, and hydrophobic core formations, have been revealed from their statistical distributions in known protein structures [16]. The inter-atomic distances are also subject to certain statistical distributions, depending on the types of the distances. Such distributions have been employed for constructing various statistical potentials for contact determination, inverse folding, structure alignment, and X-ray structure refinement [15, 17-21].

Cui et al [14] recently applied a knowledge-based approach to NMR structure refinement by extracting additional distance constraints based on the distributions of the distances in databases of known protein structures. They collected the distances of various types from structural databases and calculated the means and standard deviations of the distance distributions. They then generated the lower and upper bounds for the distances by using their means minus and plus two standard deviations (Figure 2). The results from applying these constraints to a set of NMR-determined structures showed that the structures were improved significantly after the refinement even with only a small set, restricted types of distance constraints derived from their database distributions. However, the approach has some limitations. The distance bounds derived are only simple representations of the true distributions of the distances. By restraining the distances within their bounds, the distances outside the bounds are certainly excluded, but they may still occur in real structures (although with a small probability). The distances within the bounds are also treated with equal probability, while they are actually non-uniformly distributed in most cases, and some should certainly be more preferred with a higher probability than others.

Here, we develop an alternative (or a generalized) approach of utilizing the distance distributions for structure refinement. Instead of extracting a distance range from the distribution of a distance, we use the distribution function to define a mean-force potential for the distance so that the potential is minimized when the probability of the distance in the distribution is maximized. In particular, based on the principle of statistical physics [19], we can define a so-called mean-force potential for the distance with its probability distribution (Figure 27). For a selected set of distances, we can obtain a set of mean-force potentials. The sum of the potentials can then be used to define an energy function, and a structure can be refined through energy minimization. Comparing with the approach of using the distance bounds, this approach has the advantage of being able to determine the

distances in their entire distribution ranges. The distances can also be selected more rationally based on their probability distributions. In fact, the joint probability of the distances in their distributions is maximized when the energy function defined by using the database derived mean-force potentials is minimized.

Figure 27. Typical distribution of the distance



*The distances of a specific type are typically distributed in certain range. A range constraint for the distances may be derived by restricting the distances in the most populated range, say in between mean minus and plus two standard deviations. Or, a mean-force potential may be defined for the distances based on the distribution of the distances, e.g., E = -kT ln P, where P is the distribution function, E the potential, T the temperature, and k the Boltzmann constant.

To implement the above approach, we have followed a similar procedure as used in Cui et al [14] and collected a large set of distance data from the PDB Data Bank. By using the collected data, we have calculated the distributions of the distances of different types. To facilitate the generation and analysis of the data, we have also developed a distance distribution database PIDD (Protein Inter-atomic Distance Distribution Database) for automatic processing and calculating the distances and their distributions (see [22] for detailed description of the database or check out the web server of the system at http://www.math.iastate.edu/pidd). Based on the calculated distributions, we have defined the mean-force potentials for a selected set of distances and in particular, the distances between atoms in separated residues in sequence. We then insert the potentials into the energy function of the NMR modeling software CNS (Crystallography and NMR system) [9] and used them to refine a selected set of test structures. Total 70 NMR-determined structures were refined, using the original NMR data that can be downloaded either from PDB Data Bank [2] or BioMagResBank [23]. Both original and

extended energy functions were employed. The results were compared to evaluate the effectiveness of the mean-force potentials for the refinement of the structures. Several standard measures were adopted in the comparison, including the energy values in various different categories such as the bond length energy, the bond angle energy, the NOE energy, etc., the ensemble RMSD of the structures, the RMSD of the structures against the X-ray reference structures (for available ones), and the Ramachandran plots of the structures. In terms of these measures, we have found that the structures have been improved significantly after the refinement with the database derived mean-force potentials. More specifically, we have observed significant decreases in the ensemble RMSD values and increases in the percentage of residues in the most favorable regions of the Ramachandran plots [24-25] for most of the refined structures. Of 70 tested structures, around 80% had their energy values decreased in all the categories and by 7.5% in average for overall energy. Most importantly, the NOE and dihedral angle energies were decreased substantially as well for about 65% of the structures, indicating that the mean-force potentials helped not only forming more energetically favorable structures but also forcing the structures to fit the experimental constraints even better, which was of great importance to NMR modeling [26].

The distributions of the distances

In order to estimate the distributions of the distances in known protein structures, we have downloaded 2090 X-ray crystal structures with resolution of 2.0Å or higher and sequence similarity of 70% or less from the PDB Databank. Here, we have not used NMR structures because there is an ensemble of structures for each NMR-determined protein and we need to develop an appropriate strategy for choosing the structures, which we plan to do in future. We do realize that using only X-ray crystal structures may have biases and hope that the distributions of the distances we can extract from these structures can indeed reflect common properties of the distances in all proteins. Using 70% sequence similarity cutoff is a bit arbitrary. In fact, we observe no difference in obtained distance distributions for cutoffs less than 90%. To be conservative, we then chose 70%.

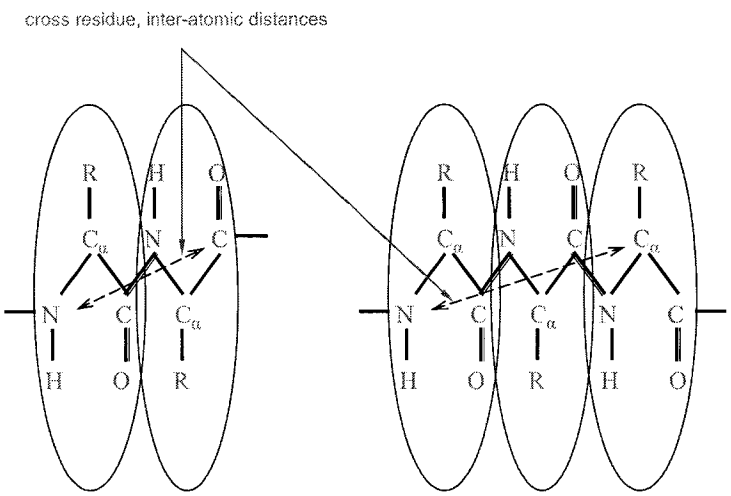
We consider the distances connecting atoms in separate residues, called cross-residue, inter-atomic distances. Such a distance can be specified by using the types of the two atoms it connects to, the types of the residues the two atoms are associated with, and the types of the residues separating the two end residues in sequence (see Figure 28). For instance, a distance between two atoms of types A_1 and A_2 contained in residues R_1 and R_2 that are separated by residues S_1, S_2, S_3 is said to have a

distance type $[A_1,A_2,R_1,R_2,S_1,S_2,S_3]$. In general, a distance between two atoms of types A_1 and A_2 contained in residues R_1 and R_2 that are separated by residues S_1, ... , S_m is said to have a distance type $[A_1,A_2,R_1,R_2,S_1,\ldots,S_m]$. Assume that the distances are distributed in a range from 0 to 20 Å. Then, for each particular type of distances, we can compute all the distances of this type from the downloaded structures and group them into a large number of, say 200, evenly divided distance intervals $[D_i, D_{i+1}]$, i = 0, 1, ... , 199 with D_0 = 0 and D_{200} = 20 Å. Let D be a distance of type $[A_1,A_2,R_1,R_2,S_1,\ldots,S_m]$. Let $P[A_1,A_2,R_1,R_2,S_1,\ldots,S_m](D)$ be the probability distribution function of D. Then, $P[A_1,A_2,R_1,R_2, S_1,\ldots,S_m](D)$ can be defined as the number of distances of type $[A_1,A_2,R_1,R_2, S_1,\ldots,S_m]$ found in $[D_i, D_{i+1}]$ divided by the number of distances of the same type found in the whole distance interval $[D_0, D_{200}]$, for any D in $[D_i, D_{i+1}]$, i = 0,1, ... , 199, i.e.,

$$P[A_1,A_2,R_1,R_2,S_1,\ldots,S_m](D)=\frac{\#D's\in[D_i,D_{i+1}]}{\#D's\in[D_0,D_{200}]},\quad D\in[D_i,D_{i+1}],\quad i=0,\ldots,199,$$

where $D's$ means the distances of type $[A_1,A_2,R_1,R_2, S_1,\ldots,S_m]$.

Figure 28. Cross residue, inter-atomic distances

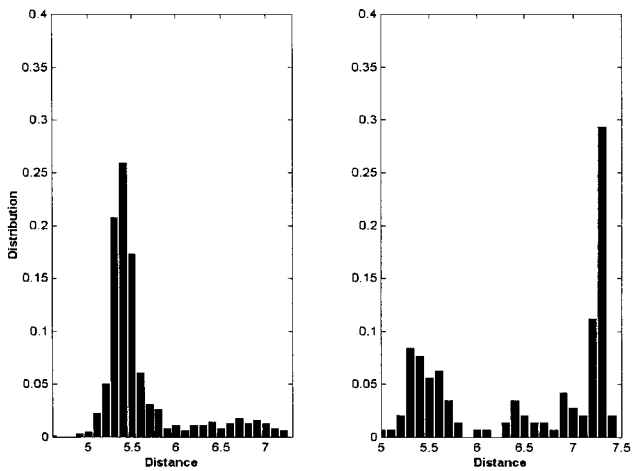


*The distances are specified by the types of the two atoms they connect to, the types of the residues the two atoms are associated with, and the types of the residues separating the two end residues in sequence.

Figure 29 shows two distance distribution functions obtained by using the above formula. The graph on the left is the distribution of the distances between C_α in TYR and C_α in TYR separated by LYS in sequence, which has a peak around D = 5.4 Å and a long tail. The graph on the right is the distribution of the distances between C_α in SER and C_α in TRP separated by GLY in sequence, which has two peaks around D = 5.4 and 7.3 Å, respectively. The graphs show clear non-uniform

distributions of the distances. It agrees with the fact that large portions of protein segments form regular secondary structures, i.e., α-helices or β-sheets, and therefore, the short-range distances are more likely to be distributed around the values that are preferred by the secondary structures. Depending on the types of the distances, they may occur more in α-helices or in β-sheets or both, and their distribution functions will accordingly have one peak around a distance value that may be preferred by α-helices or β-sheets, or two peaks around two distance values, one preferred by α-helices and another by β-sheets.

Figure 29. Samples of distance distributions*



*The graph on the left is the distribution of the distances between Cα in TYR and Cα in TYR separated by LYS in sequence. The graph on the right is the distribution of the distances between Cα in SER and Cα in TRP separated by GLY in sequence.

Let's assume that in average, there are 10 different atom types for A_1 and A_2. For each of R_1, R_2, S_1, … , S_m, there are 20 different residue types. Therefore, in total, there can be $10^2 \times 20^{m+2}$ different types of distances. Even if only three separating residues are allowed, the total number of distance types can be as many as 320,000,000. In order to collect and process the enormous amount of data, we have developed a database for automatically generating, computing, and analyzing the distributions of the distances. Currently, the database can generate the distribution data for all short-range distances with up to three separating residues. By using the database's web interface, we can select distance types of interest and automatically generate the distribution functions as shown in Figure 29 (see [22] or visit http://www.math.iastate.edu/pidd for more details).

In this work, we have only investigated short-range distance types and in particular, the distance types like $[A_1,A_2,R_1,R_2,S]$ with only one possible separating residue. Furthermore, we have simplified the distance types by using $S = 0$ or 1 to indicate if the distances are separated by zero or one residue, no matter what residue type is. In this case, the total number of possible distance types becomes $2 \times 10^2 \times 20^2 = 80{,}000$. The reason we have only considered such a set of distance types is because it is probably the simplest set of short-range distances we can start investigating with, yet is already sufficient enough to show that the constraints or potentials derived for the distances can be used to refine protein structures effectively. The distances can certainly be extended to include longer-ranges and more complicated types for possible more extensive and effective uses. It however requires more substantial work, which we plan to pursue in future.

Distance-based mean force potentials

Since the distributions of the distances are non-uniform in general, constraints on the distances can immediately be extracted based on these distributions. As we have mentioned in the introduction section, Cui et al. [14] have derived bound constraints on the distances by using the means minus and plus two standard deviations of the distances as the lower and upper bounds, and applied the constraints to the refinement of NMR-determined protein structures. The advantage of this approach is that the constraints are easy to generate and straightforward to implement with current NMR modeling software such as CNS because they can be applied for structure refinement in the same way as the NOE distance constraints. However, by using simple bounds, the information on the distances demonstrated in the distributions of the distances is not completely exploited, since the constraints exclude the possible distance values outside the bounds and also treat the distance values inside the bounds equally. In fact, the distances outside the bounds are still likely although with only small chances. Also, the distances inside the bounds are obviously distributed non-uniformly and the more probable ones should be considered with higher priorities. A relatively more complete approach is to incorporate the information in the distribution functions as much as possible to restrict the use of the distances. To this end, for each type of distance, a potential function can be defined by using the distribution function for the distance so that the potential energy is minimized when the distance maximizes the probability distribution. One of such potential function can be defined with the idea of mean-force potentials in the statistical physics [19]. Let $P_{i,j}$ be the probability distribution for any distance of interest between atoms i and j. Then, the mean-force potential $E_{i,j}$ for the distance can be defined such that for any D, $E_{i,j}(D) = -k_B \, T \ln P_{i,j}(D)$, where k_B is the Boltzmann constant and T the

temperature. Let S be a set of atom pairs selected to define the mean-force potentials. Then, the sum of the mean-force potentials for the atom pairs in S can be defined as the mean-force energy E_{PMF} for S:

$$E_{PMF} = \sum_{(i,j)\in S} E_{i,j}(\| x_i - x_j \|) = -k_B T \sum_{(i,j)\in S} \ln P_{i,j}(\| x_i - x_j \|).$$

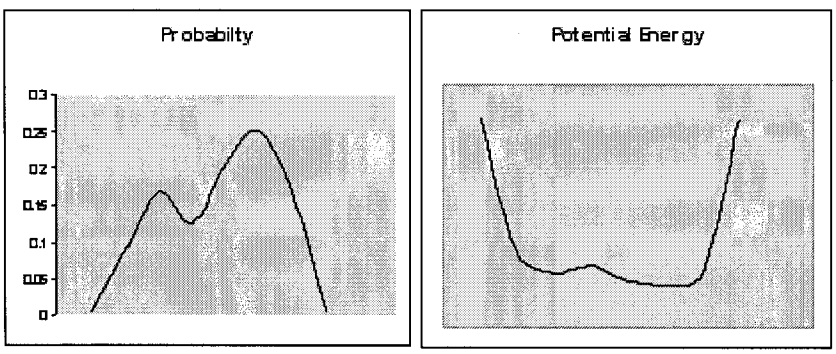
Here, the mean-force energy E_{PMF} is minimized when the joint probability of $P_{i,j}$ for all (i,j) in S is maximized. Therefore, to choose the most probable distances for a protein based on the probability distribution functions $P_{i,j}$, it is equivalent to solve an optimization problem to minimize the mean-force energy E_{PMF}. The function E_{PMF} may not necessarily be easy to minimize if a global minimum is to be found. It may become even harder when $P_{i,j}$ have multiple peaks (or equivalently, $E_{i,j}$ have multiple minima, see Figure 30).

In this work, for each structure to be refined, we first find the database distributions for the distances of $[A_1,A_2,R_1,R_2,S]$ types in the structure, where A_1 and A_2 included all backbone and side-chain atoms except for hydrogen atoms. We then approximated each distribution graph by a normal distribution function,

$$P_{ij}(D) \approx \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(D-\mu)^2}{2\sigma^2}],$$

with μ and σ determined by using least-squares interpolation. In the end, we only kept the functions that can be approximated relatively accurately and used them to obtain the mean-force potentials. In this way, the graphs with clear multiple peaks were removed from further consideration and the selected potentials had "unique" minima. Here, we sacrificed some distribution data but kept the potentials simpler and easier to optimize.

Figure 30. Mean-force potential vs. probability distribution



The mean-force potential is minimized when the corresponding probability distribution function is maximized. The mean-force potential will have multiple minima if the probability distribution function has multiple peaks.

With the above approximation, each mean-force potential becomes a quadratic function and could be obtained easily. Let E_{PMF} be the sum of all these potentials that are defined for the selected distances in the given structure. As we have mentioned, the structure can be refined by minimizing E_{PMF}, which can be done with an optimization routine implemented in protein modeling software. We have used CNS for our NMR structure refinement. In this software, in addition to routines for handling initial structure generation and experimental data processing, a particular routine is implemented to refine an NMR structure by minimizing a set of energy potentials including the bond-length and bond-angle potentials (E_{bond} and E_{angle}), the electrostatic and van der Waals potentials (E_{elec} and E_{VDW}), the improper angle potentials (E_{imp}), and the NOE and dihedral angle potentials (E_{NOE} and E_{DIH}) [9]. The last two types of potentials are used for minimizing the violation of the experimental NOE distance and dihedral constraints. In order for the selected distances in the structure to agree with their database distributions, we have added the sum of their mean-force potentials, E_{PMF}, into the CNS built-in potentials. Then, the energy function becomes

$$E = E_{bond} + E_{angle} + E_{elec} + E_{VDW} + E_{rep} + E_{NOE} + E_{DIH} + E_{PMF},$$

and the structure can hopefully be refined when this new energy function E is minimized.

Refining NMR structures

We have tested the database derived mean-force potentials for the refinement of NMR-determined structures. The original NMR experimental constraints for the structures were downloaded from PDB Data Bank and BioMagResBank. CNS was used for all the computation. Total 70 structures were selected as the test cases. The structures were selected mainly because they had the original NMR data available and the data format was acceptable by CNS. The structures were refined using the geometric embedding and energy minimization routines implemented in CNS. The results obtained with and without the database derived mean-force potentials were compared and assessed in terms of several standard measures used in NMR modeling, including the potential energy of the structures in various categories, the RMSD values of the ensembles of structures, and the RMSD values of the structures compared with their X-ray reference structures (for available ones), and the Ramachandran plots.

CNS can be used to refine either X-ray or NMR structures. The part for NMR structure refinement contains four steps: connectivity calculation, template generation, annealing, and

acceptance test. Connectivity calculation takes the protein sequence as the input and produces a connectivity file for the backbone of the protein. Template generation uses the connectivity file to construct an extended structure (or a group of extended structures) for the protein as the initial structures for annealing. The annealing process has two options, one with simple simulated annealing and another with distance geometry simulated annealing. The latter embeds the structure in 3D by satisfying the distance constraints (geometric embedding) before doing simulated annealing (energy minimization). The last step, acceptance test, evaluates the structures with a group of acceptance criteria including the satisfaction of various experimental constraints and stereochemistry requirements. In our calculations, we have used the option for distance geometry simulated annealing with the database derived mean-force potentials included in the CNS built-in energy function. Therefore, the structures were determined with geometric embedding followed by energy minimization. Typically, geometric embedding helps form a structure or an ensemble of structures that satisfy a high percentage of given distance constraints, but there may still be constraints violated. In addition, the structures may not be energetically favorable. Therefore, the following energy minimization is always necessary. Since the energy function includes the classical force field potentials and the potential terms for NMR constraints satisfaction, energy minimization not only helps minimizing the potential energy but also further reduces the violation of the NMR constraints. By including the mean-force potentials in the energy function, the structures were expected to be further refined by choosing more probable distances according to their distributions in known protein structures.

Energy of structural ensembles

The energy function in CNS includes the bond-length and bond angle potentials (E_{bond} and E_{angle}), the potentials for improper angles (E_{imp}), and the potentials due to electrostatic and Van der Waals interactions (E_{elec} and E_{VDW}). In addition, there are also terms defined for NOE distance and dihedral angle constraints (E_{NOE} and E_{DIH}). The sum of the terms measures how much the constraints are satisfied. When the energy function ($E_{Overall}$) is minimized, the structure is considered to be both energetically favorable and experimentally feasible. In other words, the lower the energy is, the better the structure is considered in terms of the intrinsic physical interaction and experimental constraint satisfaction. We have refined the selected NMR structures using the CNS distance geometry / dynamic simulated annealing protocol with original NMR experimental distance and dihedral angle constraints. We recorded the energy values of the structures in the structural ensemble for each

protein determined with and without using the database derived mean-force potentials (which we call CNS and CNS-PMF, respectively). Table 1 shows the energy values for a list of refined structures in various categories and in particular, the means and standard deviations of the energy values in each structural ensemble. Note that for a fare comparison, the calculation of the overall energy did not count the contribution from the mean-force potentials although the latter were used in the CNS+PMF refinement. Note also that the energy due to electrostatic interactions was not listed because the corresponding potentials were not included in the default CNS refinement protocol. From Table 5, we observed that the means and standard deviations of the energy values of the ensembles of structures became smaller in almost all categories after the structures were refined with the addition of the database derived mean force potentials. The results suggested that the refined structures, when using database derived mean-force potentials, were clearly more favorable energetically. Surprisingly, they also satisfied the experimental constraints better as the NOE and DIH energies were decreased in many cases as well. Overall, in terms of the means and standard deviations of the energy values in the structural ensembles, of the 70 selected NMR structures, about 80% had the overall energy significantly reduced, in average by 7.5%, and about 65% had the NOE energy decreased, in average by 5%, after refined with additional database derived mean-force potentials. Here we have not calculated the statistics for the DIH energy because some structures did not have the DIH data and energy available.

Table 5. Energy of NMR-determined ensembles after general and refined methods*

PDB	Method	Over all (kj/mol)	Bond	Angle	Impulsive	Van der Vaals	Noe	diheral
1AFI	CNS	160.9±72.0	6.2±3.3	63.6±18.8	8.4±7.2	54.2±21.7	27.6±20.1	0.9±0.9
	CNS+PMF	122.1±56.5	4.2±2.3	53.9±15.8	6.2±4.7	37.8±17.3	19.0±15.4	1.0±1.1
1BA4	CNS	93±60.8	4.0±3.0	34.3±21.8	4.4±5.9	26.0±14.3	24.3±15.9	0
	CNS+PMF	57.8±14.7	2.1±0.7	24.1±3.7	2.1±1.2	17.1±4.0	12.4±5.2	0
1DKC	CNS	155.7±90.1	7.4±4.1	40.1±10.6	4.7±2.5	48.9±48.6	54.6±24.3	0
	CNS+PMF	118.6±40.4	5.2±2.0	31.4±8.1	3.2±2.1	34.6±12.4	44.3±15.8	0
1DVV	CNS	85.6±19.6	3.1±0.9	40.7±5.8	4.0±1.1	23.7±7.8	14±5.2	0.05±0.06
	CNS+PMF	73.3±15.8	2.5±0.9	37.5±3.7	3.5±0.9	18.4±4.7	11.2±5.5	0.03±0.02
1I6F	CNS	190.0±73.2	1.4±2.1	24.4±8.8	1.3±1.9	113.8±47.3	48.9±12.9	0.16±0.47
	CNS+PMF	173.8±8.3	0.9±0.3	22.6±1.8	0.9±0.5	103.4±3.3	45.9±2.4	0.06±0.09

* Listed are means and standard deviations of the energies of the structural ensembles in various categories: Overall – total energy; Bond – bond-length energy; Angle – bond-angle energy; Improper – improper angle

energy; Van der Waals – Van der Waals interaction energy; NOE – energy for NOE distance constraint satisfaction; DIH – energy for dihedral angle constraints. CNS – refined with original NMR data and CNS built-in energy function. CNS+PMF – refined with original NMR data, CNS built-in energy function, and database derived mean-force potentials.

RMSD of structural ensembles

The precision of an ensemble of structures determined by NMR usually is measured by the RMSD values of the structures in the ensemble compared with the average structure of the ensemble, and in particular, by the mean and standard deviation of these values [8]. The precision may be overestimated since the ensemble of structures determined by current modeling software may not necessarily contain the whole range of structures determined by the given distance constraints [14]. Nevertheless, as shown in Table 6, the means and standard deviations of the RMSD values for the listed ensembles of structures all became smaller after the structures were refined with database derived mean-force potentials. Overall, of 70 selected NMR structures, about 65% had the means and standard deviations of the ensemble RMSD values reduced after refined with database derived mean-force potentials, and the ensembles of structures hence became relatively more converging or compact as can be seen from the example given in Figure 31. Without further experimental evidence, of course, it is hard to say whether the refined ensembles of structures reflected the structural fluctuations more accurately. However, the refined ensembles were indeed more compact consistently, especially in the loop regions where there were not sufficient NMR experimental constraints. Similar results were observed in other related reports [27].

Table 6. Precision of NMR-determined ensembles*

PDB	Method	RMSD(Å)
1AFI	CNS	0.89±0.26
	CNS+PMf	0.62±0.20
1BA4	CNS	4.2±1.1
	CNS+PMF	4.84±1.28
1DKC	CNS	4.3±1.17
	CNS+PMF	3.94±1.11
1DVV	CNS	1.44±0.36
	CNS+PMF	0.79±0.22
1I6F	CNS	1.27±0.42
	CNS+PMF	0.91±0.40
1JKZ	CNS	0.82±0.29
	CNS+PMF	0.98±0.44
1M94	CNS	1.08±0.28
	CNS+PMF	0.85±0.23

* Shown in the table are the means ± standard deviations of the RMSD values of the structures in the structural ensembles compared with the average structures. CNS – refined with original NMR data and CNS energy function. CNS+PMF -- refined with original NMR data, CNS energy function, and database derived mean-force potentials.

Figure 31. The superimpositions of 1I6F ensembles*

*Left: refined by regular CNS refinement protocol. Right: refined by CNS plus database derived mean-force potentials. The structures were aligned and displayed by using MolMol graphics software [28]. The ensembles appeared more compact, especially in the loop and terminal regions where there were not sufficient NMR experimental constraints.

Comparison with X-ray reference structures

Table 7. RMSD against X-ray reference structures*

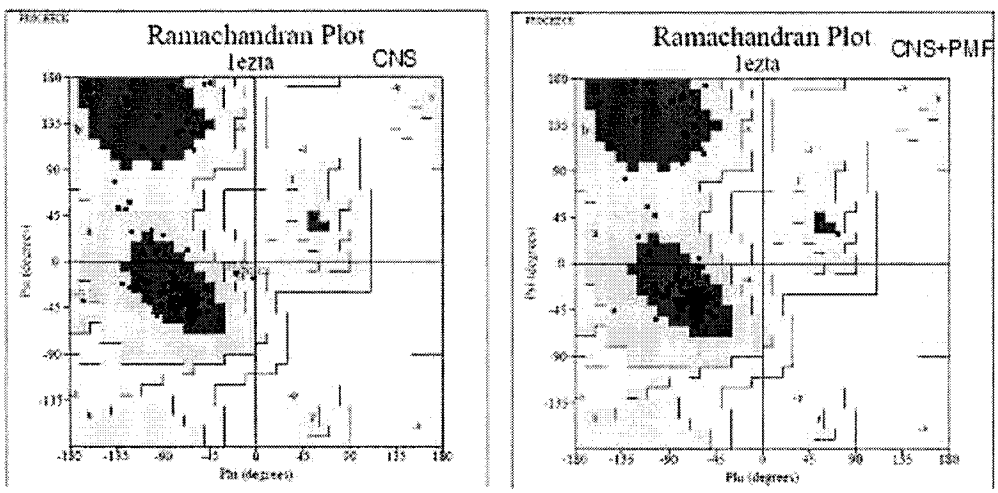
NMR	CRY	CNS	CNS+PMF
1BCN	1HIK	2.99±0.27	2.94±0.14
1CRP	121P	2.06±0.11	2.02±0.09
1E8L	193L	2.33±0.15	2.36±0.18
1GB1	1PGB	1.28±0.10	1.24±0.06
1ITL	1RCB	2.91±0.09	2.81±0.13
1JOR	1SNO	2.55±0.14	2.47±0.17
1KUN	2KNT	2.51±0.24	2.56±0.25
2IGG	1PGB	1.91±0.34	1.89±0.21
3PHY	1NWZ	3.28±0.21	3.21±0.17

*Shown in the table are the means ± standard deviations of the RMSD values of the structures in the structural ensembles compared with the X-ray reference structures. CNS – refined with original NMR data and CNS energy function. CNS+PMF – refined with original NMR data, CNS energy function, and database derived mean-force potentials.

We have also selected a small set of refined NMR structures (1BCN, 1CRP, 1E8L, 1GB1, 1ITL, 1JOR, 1KUN, 2IGG, 3PHY) and compared them with their X-ray reference structures in terms of the RMSD values of the pairs of NMR and X-ray structures. Since each protein has an ensemble of NMR structures, the mean and standard deviation of the RMSD values of the member structures were calculated and used as an assessment for the whole ensemble of structures. As shown in Table 7, in most cases, both means and standard deviations of the RMSD values for the ensembles of structures refined with additional database derived mean-force potentials were smaller than those refined without them. The differences were not so large. However, the RMSD values were average measures on overall structural differences. Therefore, the small RMSD differences between the structures refined with or without database derived mean-force potentials as shown in Table 7 may still imply large local structural differences, which can be analyzed case by case in practice.

Ramachandran plots

Figure 32. Ramachandran plots of original and refined protein structures*



*Left, the Ramachandran plot of protein structure 1EZT generated by original method (CNS). Right, the Ramachandran plot of protein structure 1EZT generated using additional mean force potential. Note that three residues in generously allowed region originally have moved to the most favorable region after employing the distance derived potentials. (By Procheck and AQUA)

To further evaluate the refined NMR structures, we have also examined the sequential φ and ψ angles for the residues in the structures using PROCHECK [25]. In particular, we have checked the Ramachandran plots for all the structures and calculated the percentages of the residues in different plot regions, most favorable, additional allowed, generously allowed, and disallowed [24]. A Ramachandran plot is a two-dimensional graph in the φ-ψ-plane. The plot has the above four major regions indicating the preferences of φ and ψ angles for protein residues. The φ-ψ angles formed in a residue can be represented by a point in the Ramachandran plot. If the point is in a particular region, we simply say that the corresponding residue is in that region. Usually a well-refined structure has a high percentage of residues in the most favorable region of the plot. We have compared the Ramachandran plots of the structures refined with and without using database derived mean-force potentials. Since each protein has an ensemble of structures, we have only compared the Ramachandran plots for the averaged and minimized structures (obtained by minimizing the energy of the proteins started with the average structures of the structural ensembles [9]). The results showed that many structures, after refined with database derived mean-force potentials, had higher percentages of residues in the most favorable regions of the Ramachandran plots. Figure 32, in particular, showed an example, where the Ramachandran plots of the averaged and minimized structures for 1EZT refined with and without using the database derived mean-force potentials are displayed. The plots showed that many residues in the generously allowed region have moved to the most favorable region after the structure was refined with the database derived mean-force potentials.

A statistical analysis showed that for the 70 NMR structures we have examined, the percentage of the residues in the most favorable region of the Ramachandran plot for each protein was increased, in average, from 69.1% to 73.4%, after the structures were first refined with the original NMR data and CNS built-in potentials and then with additional database derived mean-force potentials, while the percentage of residues in the disallowed region was decreased, in average, from 3.3% to 2.2% (see Table 8 (a), (b)). These results further demonstrated that the database derived mean-force potentials helped the structures to form more favorable local conformations even just in terms of the sequential φ and ψ angles of the protein residues.

We also applied the method to comparative models in CASPR and such improvement has also been obtained (see more details in Appendix E).

Table 8. Statistics on Ramachandran plots of selected proteins

In average*	CNS	CNS+PMF
Most favored	69.1±13.1	73.4±12.1
Additional allowed	22.0±7.9	19.2±7.9
generously allowed	5.6±5.3	5.1±4.5
disallowed	3.3±3.4	2.2±2.9

*(a) It shows the mean percentage and standard deviation of residues in each region by two different approach, the original one (CNS) and the modified one (CNS+PMF).

Percentage*	Most favored	Disallowed
Improvement	82%	47%
No change	8%	40%
Worse	10%	13%

*(b) It sows the percentage of proteins which have improvement, non-change or getting-worse on most favored and disallowed regions, using mean force potentials.

Concluding remarks

In this paper, we have investigated an alternative, generalized, and in certain sense, improved approach of utilizing the distributions of the protein inter-atomic distances in databases of known protein structures for structure refinement as proposed in Cui et al [14]. Instead of extracting the distance ranges from the distributions of the distances, here we used the distribution functions to define a set of mean-force potentials for the distances. We have applied the derived potentials for refining a set of NMR determined structures and obtained positive results in terms of several standard measures. In particular, we have observed significant decreases in the ensemble RMSD values and increases in the percentage of residues in the most favorable regions of the Ramachandran plots for most of the refined structures. Of 70 tested structures, around 80% had their energy values decreased in all the categories and by 7.5% in average for overall energy. Most importantly, the NOE and dihedral angle energies were decreased substantially as well for many cases, indicating that the mean-force potentials helped not only forming more energetically favorable structures but also forcing the structures to fit the experimental constraints even better, which was of great importance to NMR modeling.

The distance types we have examined in this work included only short range distances with up to one separating residue. While the mean-force potentials for these distance types already showed

the promising results for using the database derived mean-force potentials for NMR structure refinement in general, further extension of the work to include longer range distance types is necessary and is expected to make the whole approach to be even more powerful and effective for structure refinement. The immediate extension is perhaps to consider distance types with up to three separating residues with the types of the separating residues also specified. The distributions of such distance types have already been readily accessible through the protein inter-atomic distance distribution database PIDD developed by the authors. A database for longer range distances with more than three separating residues can also be built, probably without necessarily specifying the types of the separating residues for the reason that they may not affect the distributions of the long-range distances.

In this work, we have also selected only the distance types whose distribution functions contained single peaks since otherwise, the mean-force potentials may not be so easy to minimize. As we have mentioned in the paper, the multiple peaks in the distribution functions can well be attributed to the distributions of the distances in different types of secondary structures. So, further classification of the distance types by using say their occurrences in α-helices or β-sheets may be useful for obtaining more specific, single peak distribution functions. The more specifically the distance types can be defined, the more effectively the corresponding distance constraints or mean-force potentials can be applied to structure refinement, and of course, the more difficult to obtain sufficient distance samples as well, given the limited amount of known structures available. We plan to investigate these issues in future work.

Based on Cui et al [14] and this work, both distance range constraints and mean-force potentials have been proved to be useful for building knowledge-based refinement models for NMR-determined structures. We have not been able to make head to head comparisons between the two approaches, but the differences seemed depend on how the constraints or the potentials were selected and applied. The advantage of using the distance range constraints is that they can be extracted from the distribution functions easily and included in NMR refinement data straightforwardly. However, the bounds on the ranges exclude some possible distances and treat the distances inside the bounds equally. In this sense, the mean-force potentials may provide more complete distribution information on the distances and allow the distances selected more rationally according to their probability distributions in known protein structures. In any case, with the increasing number of high-resolution protein structures being determined, many commonly shared conformational properties such as the formation of various types of inter-atomic distances can be obtained based on the statistical distributions of the properties in databases of known protein structures. These database derived

properties can then be employed for many important modeling purposes including NMR structural refinement. Indeed, we have also been able to extend the work described in this paper to the refinement of comparative protein models in our recent participation in the 2006 CASPR comparative model refinement competition (http://www.predictioncenter.org/). Further development or expansion of the work will be made in several possible directions and be reported elsewhere.

References

1. Wuthrich, K., NMR of Proteins and Nucleic Acids, Wiley, New York, 1986
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., The Protein Data Bank. *Nucleic Acids Research*, 28 pp. 235-242, 2000.
3. Clore, G. M. and Gronenborn, A. M. New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. USA* **95**. 5891-5898 (1998).
4. Bax, A., Vuister, G.W., Grzesiek, S., Delaglio, F., Wang, A.C., Tschudin, R., Zhu, G., Measurement of homo- and heteronuclear couplings from quantitative J correlation. *Methods Enzymol* 1994, 239:79-125.
5. Abseher, R., Horstink, L., Hilbers, CW., Nilges, M.. Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins* 1998, 31: 370–382.
6. Ishima, R., Torchia, DA., Protein dynamics from NMR. *Nature Struct Biol*, 2000, 7: 740–743.
7. Doreleijers, J.F., Rulmann, J.A.C., and Katein, R., Quality assessment of NMR structures: A statistical survey, *Journal of Molecular Biology*, 281, 149-164 (1998).
8. Spronk, C.A.E.M., Natuurs, S.B., Bonvin, A.M.J.J., Krieger, E., Vuister, G.W., and Vriend, G., The precision of NMR structure ensembles revisited. *Journal of Biomolecular NMR* 25. 225-234 (2003).
9. Brunger, AT., Adams, PD., Clore, GM., DeLano, WL., Gros, P., Grosse-Kunstleve, RW., Jiang, JS., Kuszewski, J., Nilges, M., Pannu, NS., Read, RJ., Rice, LM., Simonson, T., Warren, GL., Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*. 1998 Sep 1;54 (Pt 5):905-21.
10. Tjandra, N., Omichinski, J.G., Gronenborn, A.M., Clore, G.M. and Bax, A., Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Struct. Biol.* 4, 732-738, 1994

11. Clore, G.M., Gronenborn, A.M. & Tjandra, N., Direct refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *J. Magn. Reson.* 131, 159-162. 1998

12. Kuszewski, J., Gronenborn, A. M., and Clore, G. M. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Science* **5**. 1067-1080 (1996).

13. Grishaev, A. and Bax, A. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. *J. Am. Chem. Soc.* **126**. 7281-7292 (2004).

14. Cui, F., Jernigan, R., Wu, Zj., Refinement of NMR-determined protein structures with database derived distance constraints. *J Bioinformatics and Computational Biology,* 2005, 3(6):1315-29.

15. Wall, M.E., Subramaniam, S., and Phillips, Jr.G. N. Protein Structure Determination Using a Database of Inter-Atomic Distance Probabilities. *Protein Science* **8.** 2720-2727 (1999).

16. Bourne, P. E. and Weissig, H. Structural Bioinformatics. John Wiley & Sons, Inc. 2003.

17. Miyazawa, S. and Jernigan, R. L. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18. 534-552, 1985.

18. Miyazawa, S. and Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.* 256. 623-644, 1996.

19. Sippl, M. J. Calculation of conformational ensembles from potentials of mean force, *J. Mol. Biol.* 213. 859-883, 1990.

20. Sippl, M. J. and Weitckus, S. Detection of native-like models for amino acid sequence of unknown three-dimensional structure in a database of known protein conformations. *Proteins: Structure, Function, and Genetics*, 13. 258-271, 1992.

21. Rojnuckarin, A. and Subramaniam, S. Knowledge-based potentials for protein structure. *Proteins: Structure, Function, and Genetics*, 36. 54-67, 1999.

22. Wu, D., Cui, F., Jernigan, R., and Wu, Zj., PIDD: A database for protein inter-atomic distance distribution, submitted, 2006

23. Seavey, B.R., Farr, E.A., Westler, W.M., Markley, J.L., A Relational Database for Sequence-Specific Protein NMR Data, *J. Biomolecular NMR*,1, 217-236, 1991.

24. Ramachandran, G.N., Sasiskharan, V., Conformation of polypeptides and proteins. *Advan. Prot. Chem.* 23:283-437, 1968.

25. Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R., Thornton, J.M., AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR, *J. Biomol. NMR*, 8, 477-486. 1996.

26. Nabuurs, S.B., Spronk, C.A.E.M., Vriend, G., Vuister, G.W., Concepts and tools for NMR restraint analysis and validation, *Concepts in Magnetic Resonance*, 22A (2), 90-105. 2004.

27. Cui, F., Mukhopadhyay, K., Young, W.B., Jernigan, R.L., Wu, Z., Refinement of under-Determined loops of human prion protein by database-derived distance constraints, submitted, 2006

28. Koradi, R., Billeter, M., and Wüthrich, K., MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graphics* **14**, 51-55. 1996

CHAPTER 6. LOCAL-DME CALCULATION IN PROTEIN STRUCTURE DYNAMICS

The paper to be submitted

Di Wu, Robert Jernigan and Zhijun Wu

Abstract

Flexibility and dynamics of protein structures could be captured experimentally in terms of B-factor and order parameter through Nuclear Magnetic Resonance (NMR) and X-ray Crystallography respectively. On the other hand, some theoretical approaches have been developed to predict the fluctuation of proteins in either atomic level or coarse-grained level, such as Normal Mode Analysis, Gaussian Network Model and Anisotropic Network Model. Here, we introduce a so-called Local-DME calculation, an efficient and simple analytic method to study the fluctuations of ensembles of NMR-determined protein structures. Comparison with experiments and other theoretical methods shows high correlations. Specifically, some important residues in protein function and folding can be identified.

Keywords Protein dynamics, GNM, DME

Introduction

The biological functions of proteins are highly correlated to their motions or flexibilities. Such dynamic information and fluctuations can be obtained experimentally in terms of B-factor and order parameter through Nuclear Magnetic Resonance (NMR) and X-ray Crystallography respectively [1,2]. However, experimental analysis usually provides little information regarding the ways proteins move as well as detailed dynamic information [2]. Some theoretical methods such as molecular dynamic simulation have been applied to simulate protein dynamics [3], but such all-atom detailed simulation is very expensive in computation because of complicated potential energy functions. On the other hand, some simplified methods such as Normal Mode Analysis (NMA) [4],

Gaussian Network Model (GNM) [5] and Anisotropic Network Model (ANM) [6] have also shown promising results as good as those by complicated methods in simulating protein dynamics. In general, such simplified methods involve a fewer parameters and less detailed potential energy functions, and hence are more efficient in computation, compared to the general molecular dynamic simulation.

The GNM method applies the knowledge of elastic network and Gaussian distribution to study protein motions, and only considers the residues contacts, for instance, Cα contacts. In this method, the potential energy function is dramatically simplified and contains only one single parameter, but fluctuations predicted by GNM could still have good agreement with experimental observation in fluctuation such as B-factor in crystal structures [7]. Especially, GNM method involves only one single parameter not atomic or amino acid specific. In computation, the GNM method only requires solving a singular value decomposition (SVD) problem and therefore needs much less computing than molecular dynamic simulation.

X-ray crystallography determines the unique protein structure in a crystal. For instance, the position of each atom is determined at its average position based on the electron density map, and every atom has been assigned a so-called temperature or B-factor which magnitude is proportional to the mean square displacement from its mean position. Even though such B factor values have limitations in understanding detailed atomic motions, they provide information regarding the amplitude of the fluctuations and unique source of protein dynamics in solid state experimentally. Crystal protein structures determined at the average positions of atoms are considered as equilibrium-state structures and hence could be further studied for their dynamics using theoretical approaches, which can provide the detailed information of motions and energy. On the other hand, NMR (nuclear magnetic resonance) spectroscopy provides an alternative way to determine protein structures in solution. Indeed protein structures in solution are highly related to their functions in nature, but they are also very flexible in solution and even sometimes transitions between multiple conformations could be observed through experimental data [8], all which are very crucial to understand protein functions and dynamics. However, due to insufficient experimental data from NMR experiments, structures are often underdetermined. In general, an ensemble of multiple energy-minimized structures satisfying those distance constraints instead of a unique conformation is used to represent a protein in solution. And sometimes, these models in an ensemble are deviated far from each other and being poorly determined [9-10]. Compared to crystal structures, there are not many sophisticated methods developed to theoretically study fluctuations and dynamics of NMR-determined ensembles.

Here we investigated a new computational approach to study protein dynamics of NMR ensembles in residual level (only C_α atoms). In this work, we modified DME (distance matrix error) calculations to be locally specific. For each C_α atom, only distances it involves are considered and differences of those distances between all possible pairs of two structures in an NMR ensemble are summed and represent its flexibility. Then fluctuations of each C_α atom are reproduced through this so-called Local-DME calculation and are compared with B factor values of the same protein determined by X-ray crystallography. We also apply GNM to predict the fluctuations of crystal structures as control. A detailed investigation of protein dynamics in solution and solid state is also conducted in this work.

Methods

Gaussian network model (GNM)

In GNM method, a 3D protein structure is usually described as an elastic network connected by harmonic springs with a certain cutoff distance. Only C_α atoms in each residue of a protein are considered and form an elastic network. For instance, fluctuations of C_α atoms are approximated based on Gaussian distributions of their inter C_α atomic distances around equilibrium position, and a single-parameter and non-amino-acid specific Hookean potential is adopted for the interaction. Contact matrix of C_α atoms of a protein is constructed using the Kirchhoff matrix (see equation (1)).

$$\Gamma = \begin{cases} -1 & \text{if } i \neq j \text{ and } d_{ij} \leq d_c \\ 0 & \text{if } i \neq j \text{ and } di_j \geq d_c \\ \sum_{i,i\neq j} \Gamma_{ij} & \text{if } i = j \end{cases} \qquad (1)$$

where i and j are indices of C_α atom in the sequence of a protein chain, d_{ij} is the distance between ith and jth C_α atoms, and d_c is the cut off distance, usually 7Å.

The mean-square fluctuation of each C_α atom and cross-correlation between any two C_α atoms can therefore be evaluated through the inverse of Kirchhoff matrix (see equation (2)).

$$< \Delta R_i^{\,2} >= \gamma k_B T \ [\Gamma^{-1}]_{ii}$$
$$< \Delta R_i \cdot \Delta R_j >= \gamma k_B T \ [\Gamma^{-1}]_{ij}, \quad i \neq j \quad (2)$$

where $[\Gamma^{-1}]_{ii}$ and $[\Gamma^{-1}]_{ij}$ are read from the diagonal or off diagonal of the inverse matrix of Γ^{-1} , T is the temperature, k_B is the Boltzmann constant, γ is a scaling factor and ΔR_i is the column vector of the fluctuation of the ith C_α atom.

In general, Γ is symmetric and positive semi-definite and hence the singular value decomposition could be applied to compute its identical pseudo inverse through equations (3),

$$V \Sigma V^T = \Gamma$$
$$\Gamma^{-1} = V^T \Sigma^{-1} V, \qquad (3)$$

where Σ is a diagonal matrix of Γ^{-1} , and V is singular vector matrix and orthogonal. Therefore, those mean-square fluctuation and cross-correlations could be obtained once the inverse is available. And usually only non-zero singular values as well as their corresponding singular vectors are considered.

Local-DME calculation

The difference between two conformations of the same protein could be calculated using DME (distance matrix error) method, which can give an averaged deviation between two structures considering all atoms. In DME calculations, the pair wise inter-atomic distance matrix for each conformation will be generated respectively and the Frobenius norm of difference matrix of these two matrices then could be computed to show the averaged deviations between these two structures (see equation (4)),

$$[C]_{ij} = c_{ij} = \| x_i^{\,c} - x_j^{\,c} \|_2, \ [D]_{ij} = d_{ij} = \| x_i^{\,d} - x_j^{\,d} \|_2,$$
$$DME\ (C,D) = \frac{\| C - D \|_F}{n(n-1)/2} = \frac{(\sum_{i}^{n} \sum_{j=i+1}^{n} (c_{ij} - d_{ij})^2)^{1/2}}{n(n-1)/2} \qquad (4)$$

where C is the generated distance matrix of one structure, D is the generated distance matrix of the other structure, $||\ ||_2$ is the norm and $||\ ||_F$ is the Frobenius norm.

However, such DME calculation only shows the difference in average between two structures and provides little information about flexibilities and deviations in the structures locally. For instance, some regions could be very flexible and hence have larger deviations, but the averaged deviation through DME calculation hardly explains that. Here the modified DME calculation has been developed to studying the local deviation specifically. For each atom, only distances it involves are considered and differences of those distances between all possible pairs of two structures in an ensemble are summed. Such local DME values are used to show the flexibility of that atom. Then fluctuations of each C_α atom are reproduced through this so-called Local-DME calculation (see equation (5) and (6)).

$$B_i^{mn} = \sum_j ([A^m]_{ij} - [A^n]_{ij}) \qquad (5)$$

$$B_i = \frac{\sum_m \sum_{n=m+1} B_i^{mn}}{l(l-1)/2} \qquad (6)$$

where l is the number of conformations in the ensemble, A_m, A_n are the distance matrices of two distinct mth and nth conformations in the ensemble, B_i^{mn} is the sum of differences of ith column between A_m and A_n, which represents the local deviation of ith atom in the ensemble, B_i is the averaged local deviation of ith atom in the whole ensemble.

Correlation calculation

We compute the linear correlation coefficient between the predicted fluctuations of NMR ensembles through Local-DME calculations and the experimental B factor values of same proteins determined by X-ray crystallography. Meanwhile, GNM has been applied to calculate the fluctuations of crystal structures, which is compared with Local-DME calculations as well. Simply, we can set up a least square problem and calculate the correlation coefficient r through equation (7).

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum x_i)^2}} \qquad (7)$$

where n is the number of residues in the protein, x_i is the experimental B factor value or calculated fluctuation through GNM for ith C_α atom, and y_i is the predicted fluctuation of ith C_α atom through Local-DME calculations.

Even though these two sets of values are not properly scaled, the calculation of correlation coefficient is still accurate. But the scale of the predicted fluctuations could be done by multiplying an appropriate constant, which could be determined through comparing experimental and theoretical data (see equation (8)).

$$x_i^s = x_i \frac{\sum_j y_j}{\sum_j x_j} \qquad (8)$$

where $\sum_j x_j$ and $\sum_j y_j$ are the sums of theoretical and experimental fluctuations of each C_α separately, and x_i and x_i^s are the theoretical fluctuations (GNM or Local DME) of ith C_α atom before and after scaling respectively.

Samples

In this work, a set of 16 proteins with both crystal structures and NMR-determined ensembles were downloaded from the PDB database [11]. For each protein, fluctuations predicted through Local-DME for the NMR ensemble and ones generated by using GNM for the crystal structure represent those theoretical B factor values and are scaled after comparing with the experimental B factor value. Those structures are listed in table 1.

The study of protein dynamics here was focused on the coarse-grained level, therefore C_α atoms of each residue were only considered in modeling fluctuations of protein. For those protein structures contain water molecules, small ligand or other cofactors, there is still no sophisticated method for incorporating these and hence were not considered in this work.

Computational tools

Matlab 7.0 installed in DELL computer with 3.0Ghz Pentium CPU and 2Gb memory is the main computational tool used in the research. To compute the inverse of Kirchhoff matrix, the singular value decomposition routine existing in Matlab 7.0 was directly called and generated the singular values and corresponding singular vectors. Based on the past experience [7], it was pointed

out that using a small amount of singular vectors in the ascending order of their singular values are sufficient enough, but in this paper, we still consider all nonzero singular values as well as their singular vectors (see source code in Appendix F).

Scaling of theoretically calculated fluctuations

For these theoretically calculated fluctuations through either GNM or Local-DME, we could determine the scaling constant through equations (8). And a detailed investigation and figures are also provided.

Results and discussions

Fluctuations predicted by GNM on crystal structures, experimental B factor values of these crystal structures and fluctuations computed using Local-DME on NMR-determined structure ensembles are compared. For each selected protein, both crystal structure and NMR determined structure ensemble were downloaded from PDB.

In GNM method, small singular values contribute significantly to the total fluctuation, which are corresponding to the slow motion modes, while large singular values and corresponding singular vectors are related to the fast motion modes, because of using reciprocal of these singular values. Even though a few smallest singular values are relatively more important in the calculation and are also sufficient enough to provide accurate prediction of fluctuations, but all nonzero singular values and corresponding singular vectors were still used here in representing the inverse of Kirchhoff matrix. Cut off distance used in GNM is used in this work.

An NMR-determined structure ensemble of a protein usually contains multiple models solved in NMR determination protocols, such as CNS (Crystallography and NMR system) [12], and all models are energy-minimized and satisfy experimental constraints in general. Local DME calculations will hence provide the predicted fluctuations of each NMR-determined structure ensemble based on the local dissimilarities on these models. Even though NMR structures are hardly compared to crystal structures in the resolution as well as accuracy, the models of ensembles are determined based on experimental data which contains much structural information including dynamics and flexibilities.

Table 8 shows the results of analysis of dynamics on proteins with both crystal structures and NMR determined structures. First two columns lists the PDB names of proteins determined by X-ray

crystallography and NMR spectroscopy respectively. The third column lists the number of amino acids of each protein. The last three columns contain comparison on fluctuations provided by different ways including theoretical and experimental methods. In the bottom of the table, the averaged correlation coefficients are also computed. For most proteins, the flexibilities predicted by Local-DME in NMR determined ensembles have high correlations with temperature factors of corresponding crystal structures (see LDME VS B-factor), and some proteins can even obtain correlation coefficient over than 0.8, such as 2PHY-3PHY and 4PTI-1PIT. The averaged correlation coefficient for LDME VS B-factor is 0.62, which indicates the protein dynamics in solution is quite similar to ones in solid state, especially some hot residues with large fluctuations, and hence using Local-DME calculation to predict the fluctuations of proteins in solution is reliable and Local-DME values can represent pseudo B-factor of NMR determined structures in a certain sense. As a control, we also applied GNM calculations to crystal structures to compute the fluctuations, in which the 7Å cutoff distance was used. For some proteins, high correlations between B-factor and fluctuations by GNM or between B-factor and fluctuations by Local DME were also obtained, but averaged correlation coefficients are 0.57 and 0.60 respectively, which are relatively lower than LDME VS B-factor.

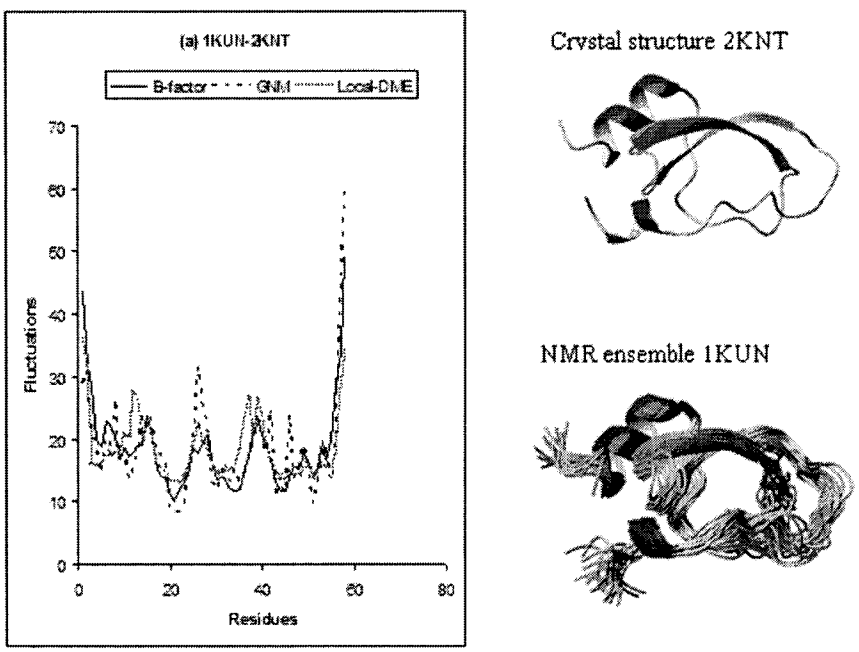
Table 9. Comparison of Local-DME and other methods in fluctuations

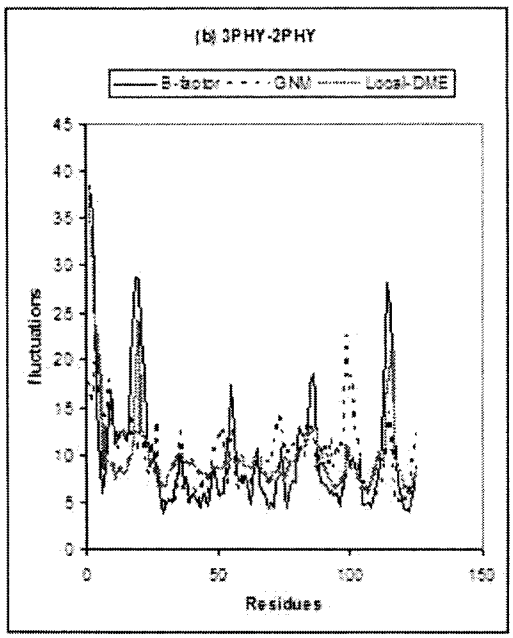
PDB Files			Correlation Coefficient*		
CRY	NMR	Residues	LDME VS B-factor	B-factor VS GNM	LDME VS GNM
2KNT	1KUN	58	0.77	0.82	0.71
1BZ6	1MYF	153	0.72	0.66	0.7
1AXJ	1FLM	122	0.58	0.65	0.83
2PHY	3PHY	125	0.82	0.53	0.48
1MBD	1MYF	153	0.57	0.58	0.63
4PTI	1PIT	58	0.85	0.74	0.78
1NOT	1XGA	13	0.69	0.52	0.53
121P	1CRP	166	0.7	0.58	0.56
1PGB	1GB1	56	0.53	0.7	0.65
1SNO	1JOR	149	0.78	0.74	0.6
1FIK	1PFL	139	0.46	0.64	0.74
1PGB	2IGG	56	0.36	0.7	0.58
1AUC	4TRX	105	0.57	0.12	0.3
9RNT	1YGW	104	0.52	0.48	0.49
1C75	1K3G	71	0.5	0.21	0.26
3EBX	1FRA	62	0.57	0.52	0.82
		mean	0.62	0.57	0.60

*It shows the comparison of experimental B factor values, fluctuations predicted by GNM on crystal structures and Local-DME values in corresponding NMR determined ensembles.

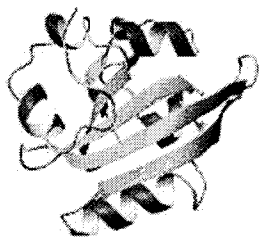
Figure 30 (1) and (2) shows the results of comparison of experimental B factor values, fluctuations predicted by GNM on crystal structures and Local-DME values in corresponding NMR determined ensembles for 2KNT-1KUN and 2PHY-3PHY, after scaling based on experimental B factor values. We also show backbone graphs of crystal structures and corresponding NMR ensembles. In 2KNT-1KUN, both Local DME values in NMR ensemble and fluctuations by GNM have high correlations with the experimental B factor values. Especially, the hot residues with large flexibilities are identified in NMR structure 1KUN as well as crystal structure 2KNT. Actually most of those hot residues are located in the surface of the protein or loop regions, and hence are relatively more flexible due to fewer contacts. In 2PHY-3PHY, the high correlation between Local DME values in NMR ensemble 3PHY and temperature factors of 2PHY was obtained, while fluctuations predicted by GNM did not give promising results, in which flexibilities of some residues were either overestimated or underestimated. Figure 30 (3) shows same comparison as (1) and (2) for 1PGB-2IGG, but in this example, the correlation between Local DME in NMR ensemble 2IGG and B factor values of 1PGB was not so good, while we did obtain a better correlation between B factor values and theoretical fluctuations by GNM.

Figure 33. Plots of fluctuations of B-factor, Local-DME, GNM*

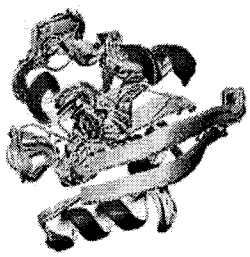


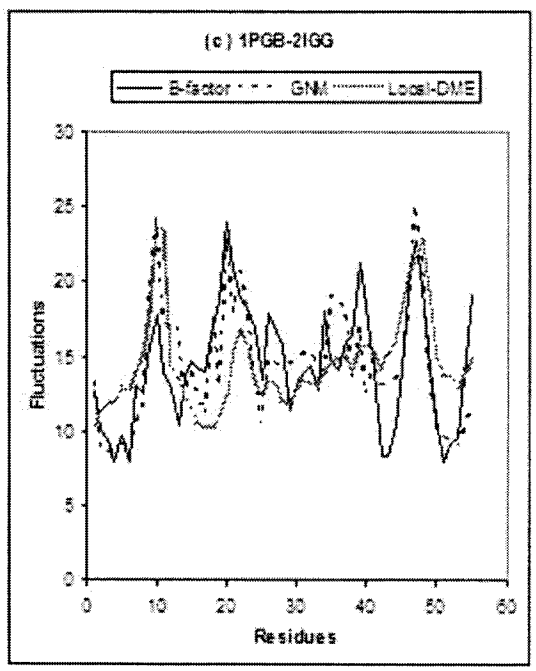


Crystal structure 2PHY

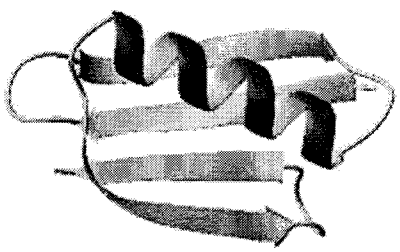


NMR ensemble 3PHY

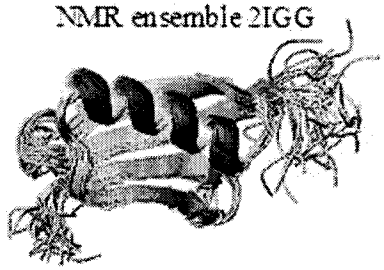




Crystal structure 1PGB



NMR ensemble 2IGG



*Plot of experimental, theoretical B factor value and Local DME against residue number. B factor value is from the experimental data in crystal structures, GNM is from the application of GNM to crystal structures and Local-DME is from the Local-DME calculations on NMR ensembles. (a) 2KNT-1KUN. Both Local-DME and GNM generated values are high correlated to B factor values. (b) 3PHY-2PHY. Local-DME calculations obtained a better correlation with the temperature factors, while GNM does not show a good result and the flexibilities in some domains were either overestimated or underestimated. (c) 2IGG-1PGB. GNM predicted the fluctuations reasonably well, while Local-DME values did not agree so well with B factor value of the crystal structure.

Conclusions and remarks

In this paper, we developed an efficient computational tool called Local-DME calculation to study the protein dynamics of NMR determined ensembles in solution. For crystal structures, it is relatively easier to obtain flexibility information of proteins in solid state through either experimental data or theoretical methods since only one structure for each protein is often determined based on the electronic density map after crystallization, while an ensemble of structures which are energy-minimized and satisfy experimental distance constraints are given in NMR structure determination instead [11]. On the other hand, many structures in an NMR ensemble are actually far deviated from the true structures and poorly being determined, which also results in difficulty in studying NMR structures theoretically, and experimental data from NMR spectroscopy is very complicated and insufficient in understanding their dynamics completely. Compared to crystal structures, there are not many theoretical approaches available to study protein dynamics of NMR structures.

In fact, an ensemble of NMR structures is determined exactly based on the experimental data which include both structural and dynamic information, and the superimposition of structures can visibly provide fluctuations of a protein in solution by some graphing software. The superimposition of structures generally requires RMSD calculations which can also provide rms values for each atom after alignment to indicate the deviations from the mean position. Hence rms values sometimes imply the flexibilities of atoms in NMR determined structures. However, the strategy of doing multiple structural alignments in an ensemble could affect results a lot and it is still very expensive in computation to find the optimal alignment using currently available techniques. Hence Local-DME provides a more accurate and reliable tool which does not require aligning multiple structures, instead, only considering the inter-atomic distances among an ensemble of structures. For the most flexible residues, the inter-atomic distances are also found to change much in different conformations of an ensemble, especially those residues in the loop or surface regions. From our results of using this

method, the hot residues which are relative more flexible in crystal structures have also been identified in corresponding NMR determined ensembles, which implies a great agreement in protein dynamics of proteins having both structures determined by NMR spectroscopy and X-ray crystallography. In our testing problems, averagely the correlation coefficient between B factor values of crystal structures and Local-DME values of NMR ensembles is 0.62, even higher than using GNM. Therefore, Local-DME calculations indeed are applicable to studying the protein dynamics of NMR structures. It is possible to use Local-DME values as pseudo B-factor for NMR structures, and the further research is on the way. However, the computational methods used in NMR structure determination and the availability of the experimental data are essential to generate NMR ensembles and hence might affect the Local-DME calculations, which could provide inconsistent information on protein dynamics. On the other hand, this can also be used to justify the quality of NMR determined ensembles. Such interesting investigation will be discussed in the future.

Acknowledgements

The support for the first author from the ISU Graduate Program on Bioinformatics and Computational Biology and ISU department of mathematics is gratefully acknowledged.

References

1. Wuthrich K.(1986) NMR of Proteins and Nucleic Acids (Wiley, New York)
2. Karplus M, McCammon JA. The internal dynamics of globular proteins. *CRC Crit Rev Biochem.* 1981;**9**(4):293–349.
3. McCammon JA, Wolynes PG, Karplus M. Picosecond dynamics of tyrosine side chains in proteins. *Biochemistry.* 1979 Mar 20;18(6):927-42.
4. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol.* 1985 Feb 5;181(3):423-47.
5. Haliloglu, T. Bahar, I., Erman Gaussian dynamics of folded proteins.,B. *Phys. Rev. Lett.* 79, 3090-3093, 1997.
6. Atilgan R, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* 2001 80:505-515.
7. Kundu S, Melton J, Sorensen D, and Phillips Jr G. Dynamics of Proteins in Crystals: Comparison of Experiment with Simple Models. *Biophys J*, August 2002, p. 723-732, Vol. 83,

No. 2

8. Wagner G, Hyberts S and Havel T. NMR STRUCTURE DETERMINATION IN SOLUTION: A Critique and Comparison with X-Ray Crystallography. *Annu.Rev.Biophys.Biomol.Struct.* 1992.21:167-98
9. Cui F, Jernigan R, Wu Z, Refinement of NMR-determined protein structures with database derived distance constraints. *J Bioinformatics and Computational Biology,* 2005, 3(6):1315-29.
10. Doreleijers J, Rulmann J, and Katein R, Quality assessment of NMR structures: A statistical survey, *J Molecular Biology*, 281, 149-164 (1998).
11. Brunger A, Adams P, Clore G, DeLano W, Gros P, Grosse-Kunstleve R, Jiang J, Kuszewski J, Nilges M, Pannu N, Read R, Rice L, Simonson T, Warren G, Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr. 1998 Sep 1;54 (Pt 5):905-21.*
12. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, 28 pp. 235-242 (2000).

CHAPTER 7. GENERAL CONCLUSIONS

General conclusions and future plans

Distance-based protein structure modeling arises from the study of protein structure modeling with the knowledge of inter-atomic distances. The challenges in this field include protein structure determination and refinement and analysis of protein structure dynamics. We introduced several algorithms and tools, which potentially have applications in related research fields.

In general, a protein structure can be determined by solving a distance geometry problem with a set of distances. The molecular distance geometry problem we studied in our research considers only sparse but exact distance data, and the main method we have developed is focused on geometric build-up algorithms. However, a general geometric build-up algorithm can be numerically unstable for some cases when the numerical errors are accumulated in a long sequence of coordinate calculations. Also, the requirement for four base atoms for the unique determination of each atom is sufficient, but not necessary, and is even redundant for rigid determination. In this work, we developed an updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse but exact distance data. The idea for the updated algorithm is based on the fact that the coordinates of any four atoms can be determined without any other information as long as all distances among them are given. On the other hand, for sparse distance data in which the general geometric build-up algorithm may fail, we incorporated the idea of rigidity and developed a rigid geometric build-up algorithm. Multiple conformations for each protein are expected to be determined and satisfying given sparse distances rather than a unique structure. The key point in this method is that the number of base atoms can be reduced from four to three so that the atom could be still determined with finite positions. Numerical testing results show that both algorithms are stable and applicable to very sparse distance data. However, we are still left with the problem of investigating the sufficient and necessary condition for protein structure modeling. The solution to this problem can help answer what is the minimum requirement of distances in modeling. When using the rigid geometric build-up algorithm, for a large system with many atoms and very sparse distance data, the number of possible conformations or substructures can be very huge and problematic, as has been seen in our testing problems. Hence, it still requires further study and additional techniques, particularly for data storage of huge combinations. Some testing results numerically blew up during the numerical computing, due to the large number of combinations even with rigid determinations of some atoms. Therefore, further investigation is still required. Another important extension of the

geometric build-up algorithms is studying the distance geometry problem with sparse but inexact distances having lower bound and upper bounds.

Due to insufficient distance data such as nuclear overhauser effect (NOE) data in NMR, the protein structures determined by conventional techniques usually are not as accurate as desired. In practice, a lot of errors could exist in given distances and some protein structures are always underdetermined. Therefore, the uses of such protein structures in important applications including homology modeling and rational drug design have been limited. We developed a novel statistical method to refine protein structures, including constructing a structural database (PIDD) and deriving so-called mean force potentials from the statistical analysis on inter-atomic distance distributions to refine NMR determined structures. First, we provided a database and structural bioinformatics system, PIDD, for distance-based protein modeling. This system can host and analyze the statistical data for protein inter-atomic distances based on their distributions in databases of known protein structures. In general, it can be used to extract geometric restraints or mean-force potentials for protein structure determination including NMR structure determination and comparative model refinement. Also, we provide a user friendly web interface so that users can easily specify the distance types and ranges, and retrieve, visualize, or download the distributions of the distances as they desire. The important application of PIDD in this work is deriving mean force potentials for NMR protein structure refinement. It is simply based on the assumption that the nature chooses the most preferred conformation because it is stable and has lowest energy. Our results show that protein structures are indeed refined in terms of energy, precision and Ramachandran plots. Such conclusions are carefully drawn from statistical analyses on a set of protein structures. However, in this method, we currently only consider the distances with one separating residue or less, and also only Gaussian-like distance distributions are considered. Therefore, there is still much room to incorporate all other possible distances, such as two-peak Gaussian distributions. Such improvements on our distance database, PIDD, as well as potential energy function, will be investigated in the future. Specifically, the current version of PIDD has provided the basic functions for processing the data for protein distance distributions. More tools will be developed to facilitate various structural analysis tasks, including tools for computing the distributions of the distances under more structural conditions, such as the distributions of the distances of certain types when they are in alpha helices vesus beta sheets. In the future, we will extend our work on PIDD to the development of a general protein geometry database that includes the statistical distribution data for other protein geometric properties besides the distances, such as all the related applications including protein structure refinement on NMR data

spectroscopy and comparative models, will be further investigated after PIDD is modified to provide enhanced and more complete analysis functions.

Finally, we have proposed an efficient computational tool, Local-DME calculation, to study the protein dynamics of NMR determined structure ensembles in solution. Local-DME provides an accurate and reliable tool which does not require aligning multiple structures. Instead, only the inter-atomic distances among an ensemble of structures is considered. For the most flexible residues, the inter-atomic distances are also found to change significantly in different conformations of an ensemble, especially those residues in the loop or surface regions. From our results of using this method, the "hot" residues which are relatively more flexible in crystal structures have also been identified in corresponding NMR determined ensembles. This implies close agreement in protein dynamics of proteins having structures determined both by NMR spectroscopy and X-ray crystallography. In our testing problems, on average, the correlation coefficient between B factor values of crystal structures and Local-DME values of NMR ensembles is 0.62, even higher than using GNM. Therefore, Local-DME calculations indeed are applicable to studying the protein dynamics of NMR structures. Specifically, it is possible to use Local-DME values as pseudo B-factors for NMR structures, and further research on this topic is ongoing.

APPENDIX A. MATLAB CODE OF GEOMETRIC BUILD-UP ALGORITHM

This is the part of source code used in geometric build-up algorithm. The first subroutine is the rigid determination. The second one is the updated geometric build-up. The last one is the general geometric build-up. The main function is not shown here.

If you need the complete code, please write to the author Di Wu(diwu@iastate.edu).

```
function [X1,x]=getco(X1,Y,d,k)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% regid and regid update function        %
% X1, base atom, Y distance for i and X1 %
% d distance matrix for X1               %
% If k=0, do not update. If k=1, update  %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if k==1
  X(1,1:3)=0;
  X(2,1) = d(2,1);
  X(2,2:3)=0;
  X(3,1) = ((d(3,1).^2) - (d(3,2).^2))/(2*X(2,1))+X(2,1)/2;
  X(3,2) = sqrt((d(3,1).^2)-(X(3,1).^2));
  X(3,3)=0;
else
   d=zeros(3,3);
   d(2,1)=norm(X1(2,:)-X1(1,:));d(3,1)=norm(X1(3,:)-X1(1,:));d(3,2)=norm(X1(3,:)-X1(2,:));
   d=d+d';
   X(1,1:3)=0;
   X(2,1) = d(2,1);
   X(2,2:3)=0;
   X(3,1) = ((d(3,1).^2) - (d(3,2).^2))/(2*X(2,1))+X(2,1)/2;
   X(3,2) = sqrt((d(3,1).^2)-(X(3,1).^2));
   X(3,3)=0;
```

```
end
x(1,1)=(Y(1)^2-Y(2)^2+X(2,1)^2)/(2*X(2,1));
x(1,2)=Y(1)^2-Y(3)^2+X(3,1)^2+X(3,2)^2-2*x(1,1)*X(3,1);x(1,2)=x(1,2)/(2*X(3,2));
x(1,3)=sqrt(Y(1)^2-x(1,1)^2-x(1,2)^2);

x(2,1)=x(1,1);
x(2,2)=x(1,2);
x(2,3)=-x(1,3);



xc=sum(X)/3;xc1=sum(X1)/3;
XX1(:,1)=X1(:,1)-xc1(1);XX1(:,2)=X1(:,2)-xc1(2);XX1(:,3)=X1(:,3)-xc1(3);
XX(:,1)=X(:,1)-xc(1);XX(:,2)=X(:,2)-xc(2);XX(:,3)=X(:,3)-xc(3);
C = XX'*XX1;
[U, S, V] = svd ( C );
Q = U * V';
x(1,1:3)=(x(1,1:3)-xc(1,1:3))*Q+xc1(1,1:3);
x(2,1:3)=(x(2,1:3)-xc(1,1:3))*Q+xc1(1,1:3);
X1(1,1:3)=XX(1,1:3)*Q+xc1(1,1:3);
X1(2,1:3)=XX(2,1:3)*Q+xc1(1,1:3);
X1(3,1:3)=XX(3,1:3)*Q+xc1(1,1:3);

----------------------------------------
function [X0,x]=build_4u(X,Y,d)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% unique update function      %
% X base atoms, Y distances   %
% d distance matrix for X     %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

X1(1,1:3)=0;
X1(2,1) = d(2,1);
X1(2,2:3)=0;
```

```
X1(3,1) = ((d(3,1).^2) - (d(3,2).^2))/(2*X1(2,1))+X1(2,1)/2;
X1(3,2) = sqrt((d(3,1).^2)-(X1(3,1).^2));
X1(3,3)=0;
X1(4,1)= ((d(4,1).^2)-(d(4,2).^2))/(2*X1(2,1))+X1(2,1)/2;
X1(4,2)= ((d(4,2).^2)-(d(4,3).^2)-((X1(4,1)-X1(2,1)).^2)+((X1(4,1)-
X1(3,1)).^2))/(2*X1(3,2))+X1(3,2)/2;
X1(4,3)= sqrt((d(4,1).^2)-(X1(4,1).^2)-(X1(4,2).^2));

A=[X1(1,1)-X1(2,1),X1(1,2)-X1(2,2),X1(1,3)-X1(2,3)
  X1(1,1)-X1(3,1),X1(1,2)-X1(3,2),X1(1,3)-X1(3,3)
  X1(1,1)-X1(4,1),X1(1,2)-X1(4,2),X1(1,3)-X1(4,3)];
A=A*2;

B=[norm(X1(1,1:3))^2-norm(X1(2,1:3))^2-(Y(1)^2-Y(2)^2)
  norm(X1(1,1:3))^2-norm(X1(3,1:3))^2-(Y(1)^2-Y(3)^2)
  norm(X1(1,1:3))^2-norm(X1(4,1:3))^2-(Y(1)^2-Y(4)^2)];
x=A\B;x=x';

xc1=sum(X1)/4;xc=sum(X)/4;
XX1(:,1)=X1(:,1)-xc1(1);XX1(:,2)=X1(:,2)-xc1(2);XX1(:,3)=X1(:,3)-xc1(3);
XX(:,1)=X(:,1)-xc(1);XX(:,2)=X(:,2)-xc(2);XX(:,3)=X(:,3)-xc(3);

C = XX1'*XX;
[U, S, V] = svd ( C );
Q = U * V';
x(1,1:3)=(x-xc1(1,1:3))*Q+xc(1,1:3);

X0(1,1:3)=XX1(1,1:3)*Q+xc(1,1:3);
X0(2,1:3)=XX1(2,1:3)*Q+xc(1,1:3);
X0(3,1:3)=XX1(3,1:3)*Q+xc(1,1:3);
X0(4,1:3)=XX1(4,1:3)*Q+xc(1,1:3);

-------------------------------------
```

```
function x=build_4(X0,Y);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% unique function              %
% X base atoms, Y distances    %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
A=[X0(1,1)-X0(2,1),X0(1,2)-X0(2,2),X0(1,3)-X0(2,3)
  X0(1,1)-X0(3,1),X0(1,2)-X0(3,2),X0(1,3)-X0(3,3)
  X0(1,1)-X0(4,1),X0(1,2)-X0(4,2),X0(1,3)-X0(4,3)];
A=A*2;

B=[norm(X0(1,1:3))^2-norm(X0(2,1:3))^2-(Y(1)^2-Y(2)^2)
  norm(X0(1,1:3))^2-norm(X0(3,1:3))^2-(Y(1)^2-Y(3)^2)
  norm(X0(1,1:3))^2-norm(X0(4,1:3))^2-(Y(1)^2-Y(4)^2)];
x=A\B;x=x';
```

APPENDIX B. INTERFACE OF THE DATABASE PIDD WRITTEN IN PERL (INCLUDING CGI, DBI, MYSQL)

Here we only list some important subroutines of the main script.

"search_data" is the distance search routine.

"graph_display" is for graphing the distance distributions.

"display_search_form" is the form providing search selections of distances

If you need the complete code, please write to the author Di Wu(diwu@iastate.edu).

```
--------------------
sub search_data
{
my ($r1,$r2,$atom1,$atom2,$s1,$s2,$s3,$sn1,$exit_not,$dp)=@_;

my @aa_data;
my ($dbh);
my ($dsn) = "DBI:mysql:proteintest:localhost";
my (%attr) = ( RaiseError => 1 );

if ($dp==1)
  { print ul(li(qq(<p style="text-align:left;"><font size="3"><font color="#FF3333">Sorry, the
distance distribution is not available since this distance type does not exist in the current PIDD
structural database!</font></p>\n)));
  }

else
{
$dbh =DBI->connect($dsn,"paddb","paddb",\%attr);
my($sth,$stmt,$count);
   if ($sn1==0)
   {$stmt=qq\select * from protein_d where a1= ? and a2 =? and atom1 = ? and atom2 =?\;
   $sth=$dbh->prepare($stmt);
   $sth->execute("$r1","$r2","$atom1","$atom2");}
```

```
elsif ($sn1==1)
{$stmt=qq\select * from protein_d where a1= ? and a2 =? and sa1=? and atom1 = ? and atom2 =?\;
$sth=$dbh->prepare($stmt);
$sth->execute("$r1","$r2","$s1","$atom1","$atom2");}
elsif ($sn1==2)
{$stmt=qq\select * from protein_d where a1= ? and a2 =? and sa1=? and sa2=? and atom1 = ? and
atom2 =?\;
$sth=$dbh->prepare($stmt);
$sth->execute("$r1","$r2","$s1","$s2","$atom1","$atom2");}
elsif ($sn1==3)
{$stmt=qq\select * from protein_d where a1= ? and a2 =? and sa1=? and sa2=? and sa3=? and
atom1 = ? and atom2 =?\;
$sth=$dbh->prepare($stmt);
$sth->execute("$r1","$r2","$s1","$s2","$s3","$atom1","$atom2");}

my @data=$sth->fetchrow_array();
if (defined ($data[0]))
  { #print $dp;
   if (($exit_not==0)&&($dp==0))
   {print ul(li(qq(<p style="text-align:left;"><font size="3"><font color="#FF3333">This
distance type is pre-queried by others.</font></p>\n)));}

   graph_play(@data);
  }
else {
     my $pp=&discovernew($r1,$r2,$atom1,$atom2,$sn1,$s1,$s2,$s3,$dp);

     $dp=$pp;
     #print $dp;
     if ($dp==2)
     {print ul(li(qq(<p style="text-align:left;"><font size="3"><font color="#FF3333">This
distance type is new and just has been studied!</font></p>)));
     my $exit_not=1;
```

```
        }
        search_data($r1,$r2,$atom1,$atom2,$s1,$s2,$s3,$sn1,$exit_not,$dp);
       }

$sth->finish();
$dbh->disconnect();
}
}

------------------------
sub graph_play
{my (@odata)=@_;

 my (@y,@x);my $sum=0;
my ($n,$m);

for (my $i=0;$i<=307;$i++)
{  if ($odata[$i+7]!=0)
   {$m=$i;}
}

for (my $i=300;$i>=0;$i--)
{  if ($odata[$i+7]!=0)
   {$n=$i;}
}

my $number=0;

my $ymax=0;


for (my $i=$n-2;$i<=$m+2;$i++)
{  #if ($odata[$i+7]>=2)
```

```
  {$y[$number]=$odata[$i+7];
  $sum+=$y[$number];
  $x[$number]=$i/10;
  $number++;
  }
}

for (my $i=0;$i<$number;$i++)

{  $y[$i]/=$sum;

  if ($ymax<$y[$i])
   {$ymax=$y[$i];}
}
my @z;

use GD::Graph::bars;

print qq(<p style="text-align:left;"><font size="3"><font color="#0000FF">This distance
distribution has total $sum sampling distances from structural database of PIDD.</font></p>\n),hr();

use constant TITLE => "Protein Atomic Distance Distribution";

my $width=500;
if (round_ceil($number/10)<4)
{ $width=round_ceil($number/10)*400;
}
elsif (round_ceil($number/10)<10)
{$width=round_ceil($number/10)*300;}

my $graph = new GD::Graph::bars($width, 300 );
my @data  = (
   [ @x ],
```

```
   [ @y ],
);


$ymax=0.25*round_ceil($ymax/0.25);

$graph->set(
   title          => TITLE,
   x_label        => "Angstrom",
   y_label        => "Distribution",
   long_ticks     => 1,
   y_max_value    => $ymax,
   y_min_value    => 0,
   y_tick_number  => 4,
   y_label_skip   => 1,
   bar_spacing    => 4,
   accent_treshold => 40,
   transparent    => 0,
   bgclr => "white",
   fgclr =>"black",
   dclrs =>['dblue'],
   accentclr =>'dblue',
);

  $graph->set_legend( "Probability" );
  my $p_number=rand(1);
  my $gd_image = $graph->plot( \@data );
  my $temp1=">/var/www/html/temp/file".$p_number.".png";
  my $temp1t=">/var/www/html/temp/file".$p_number.".txt";
  my $temp2="/temp/file"."$p_number".".png";
  my $temp2t="/temp/file"."$p_number".".txt";
```

```
 open(IMG, $temp1);
 open DIST, ">$temp1t";

printf DIST ("Distance(A)\tDistribution\n");
for (my $i=0;$i<$number;$i++)

{  if (0!=$y[$i])
   {printf DIST ("%8.2f\t%10.8f\n",$x[$i],$y[$i]);
    #print $x[$i];
   }
}
close DIST;

 print qq(<p style="text-align:left;"><font size="3"><font color="#FF0000">);
 print h4("To obtain the distribution data, click");
 print qq(<a target="_new" href=$temp2t><img border="0" src="/image/download_dist.bmp">);
 print qq(</font></p></a>\n),hr();

 print qq(<p style="text-align:left;"><font size="3"><font color="#FF0000">);
 print h4("Click to enlarge the size of the picture");
 print qq(</font></p>\n);


 binmode IMG;
 print IMG $gd_image->png;
 print qq(<a target="_new" href=$temp2><img border="0" src=$temp2 width="70%"
height="70%">);

}

----------
sub display_search_form
{my $b=shift;
```

```
print start_form(-action=>url());
 if ($b==1)
 { print qq(<p style="text-align:left;"><font size="3"><font
color="#FF3333">Warning:</font></p>\n);
   print ul(li(qq(<p style="text-align:left;"><font size="3"><font color="#FF3333">You need to
complete the form!!</font></p>\n)));}

print qq(<p style="text-align:left;"><font size="3"><font color="#0000FF">);
print h4("Step1: Select types of two end amino acids and separating amino acids (You could choose 0
to 3):");

print ("--N terminal---");
print (
    popup_menu(-name=>"pop1",
          -
values=>["X","A","R","N","D","C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V"],
          -labels=>{"X"=>"-None","A"=>"ALA A","R"=>"ARG R","N"=>"ASN N","D"=>"ASP
D","C"=>"CYS C","Q"=>"GLN Q","E"=>"GLU E","G"=>"GLY G","H"=>"HIS H","I"=>"ILE
I","L"=>"LEU L","K"=>"LYS K","M"=>"MET M","F"=>"PHE F","P"=>"PRO P","S"=>"SER
S","T"=>"THR T","W"=>"TRP W","Y"=>"TYR Y","V"=>"VAL V"},
          -default=>"-None",
          -override=>1),"--separating residues--",

  popup_menu(-name=>"pop2",
          -
values=>["X","A","R","N","D","C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V"],
          -labels=>{"X"=>"-None","A"=>"ALA A","R"=>"ARG R","N"=>"ASN N","D"=>"ASP
D","C"=>"CYS C","Q"=>"GLN Q","E"=>"GLU E","G"=>"GLY G","H"=>"HIS H","I"=>"ILE
I","L"=>"LEU L","K"=>"LYS K","M"=>"MET M","F"=>"PHE F","P"=>"PRO P","S"=>"SER
S","T"=>"THR T","W"=>"TRP W","Y"=>"TYR Y","V"=>"VAL V"},
          -default=>"-None",
          -override=>1),"---","C terminal");
```

```
print qq(</font></p>);

print qq(<p style="text-align:left;"><font size="3"><font color="#0000FF">);

print  ("# separating residues",
          radio_group(-name=>"sn1",
          -values=>["A","B","C","D"],
          -labels=>{"A"=>"None","B"=>"1","C"=>"2","D"=>"3"},
          -default=>"A",
          -override=>1));
print qq(</font></p>\n);

print   br(),
   submit(-name=>"choice",-value=>"Go To Next"),
   reset("Reset"),
   end_form();
}

sub display_search_form2
{
 my ($aa_ref,$atom_ref,$aa1,$aa2,$sn1,$t2)=@_;
 my @sn=qw\A B C D\;

 for (my $i=0;$i<4;$i++)
 { if ($sn[$i] eq $sn1)
   { $sn1=$i;
     last;
   }
}


 my @atoms=qw\A R N D C Q E G H I L K M F P S T W Y V\;
 my $k2;my $k1; my $i;
```

```
my $lab1;my $lab2;
for ($i=0;$i<=19;$i++)
 { if ($aa1 eq $atoms[$i])
    {$k1=$i+1;
    }
  if ($aa2 eq $atoms[$i])
    {$k2=$i+1;
    }
 }
 $i=0;
 foreach my $f (@{$atom_ref})
 {   $i++;
   if ($i==$k1)
     {$lab1=\%{$f};
     }
   if ($i==$k2)
     {$lab2=\%{$f};
     }
 }


 if ($aa1 eq 'X'||$aa2 eq 'X')
    {  #print $aa1;
       #print $aa2;
       display_search_form(1);
     }
  else{# my $url=url()."?aa1=".escapeHTML($aa1).".aa2=".escapeHTML($aa2);
      if ($t2)
      {print qq(<p style="text-align:left;"><font size="3"><font
color="#FF3333">Warning:</font></p>\n);
       print ul(li(qq(<p style="text-align:left;"><font size="3"><font color="#FF3333">You need to
complete the form</font></p>\n)));}
```

```
    print start_form(-action=>url());
    print qq(<p style="text-align:left;"><font size="3"><font color="#0000FF">);
    print h4("Step1: Select types of two end anmino acids:");
    print p ("Amino Acid 1:$aa_ref->{$aa1}");
    print p ("Amino Acid 2:$aa_ref->{$aa2}");
    print p ("$sn1 separating residues");
    #print p %{$lab2};
    print hr();

    print h4("Step2: Select types of $sn1 separating residues and atom for each end amino acid:");

print ("---N terminal---$aa_ref->{$aa1}-");

if ($sn1 ne 0)
 {print ("-");}

if ($sn1 ne 0)
  {print ("-$sn1 Separating Residues-");
  }

if ($sn1 ne 0)
 {print ("-");}

print ("-$aa_ref->{$aa2}---C terminal---"),

br( ),br( );
print qq(</font></p>\n);
print
  popup_menu(-name=>"atom1",
          -values=>[sort keys %{$lab1}],
          -labels=>$lab1,
          -default=>"-None",
          -override=>1);
```

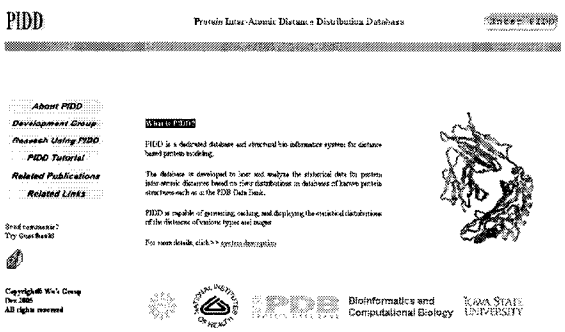
```
for ($i=0;$i<$sn1;$i++)
{ print
    popup_menu(-name=>"s".$i,
          -
values=>["X","A","R","N","D","C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V"],
          -labels=>{"X"=>"-None","A"=>"ALA A","R"=>"ARG R","N"=>"ASN N","D"=>"ASP
D","C"=>"CYS C","Q"=>"GLN Q","E"=>"GLU E","G"=>"GLY G","H"=>"HIS H","I"=>"ILE
I","L"=>"LEU L","K"=>"LYS K","M"=>"MET M","F"=>"PHE F","P"=>"PRO P","S"=>"SER
S","T"=>"THR T","W"=>"TRP W","Y"=>"TYR Y","V"=>"VAL V"},
           -default=>"-None",
           -override=>1);
}

print
   popup_menu(-name=>"atom2",
          -values=>[sort keys %{$lab2}],
          -labels=>$lab2,
           -default=>"-None",
           -override=>1),

   br(),br(),
   submit(-name=>"choice",-value=>"Submit"),
   reset("Reset"),
   end_form();
  }
}
```

APPENDIX C. TUTORIAL OF PIDD

This is the main tutorial and help file for PIDD: Database for Protein Inter-Atomic Distance Distribution. Here we only show one example. Please refer to http://www.math.iastate.edu/pidd for detailed information.

PIDD Overview

PIDD is a dedicated database and structural bio-informatics system for distance based protein modeling. The database is developed to host and analyze the statistical data for protein inter-atomic distances based on their distributions in databases of known protein structures such as in the PDB Data Bank. PIDD is capable of generating, caching, and displaying the statistical distributions of the distances of various types and ranges. The collected information can be used to extract geometric restraints or mean-force potentials for protein structure determination including NMR structure determination and comparative model refinement.



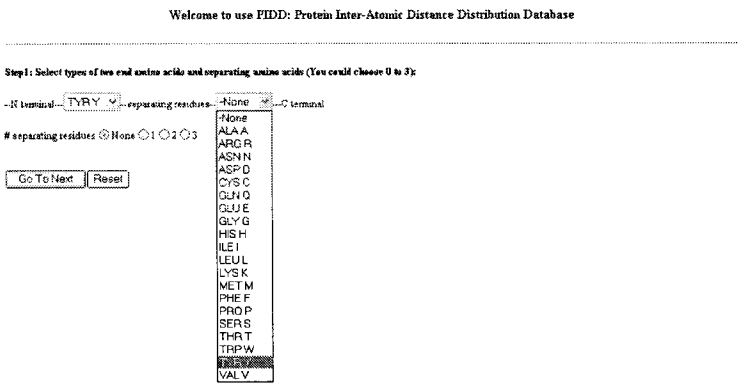
Back to Top

Example 1

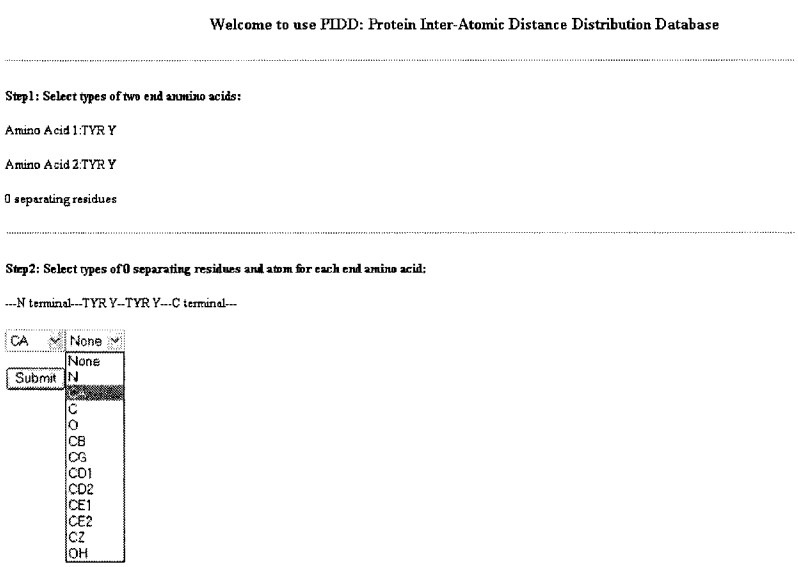
Pair wise distance distribution of atoms in two adjacent residues respectively with separating residues.

Cα in TYR and Cβ in TYR without separating residues

1. Enter the database webpage in the PIDD; Specify types of two residues where two atoms are located in respectively, and also set separating resides at None; Click Go To Next.



2. Specify types of two atoms ; Click Submit.



3. Display the results

Step3: Display Searching Results.

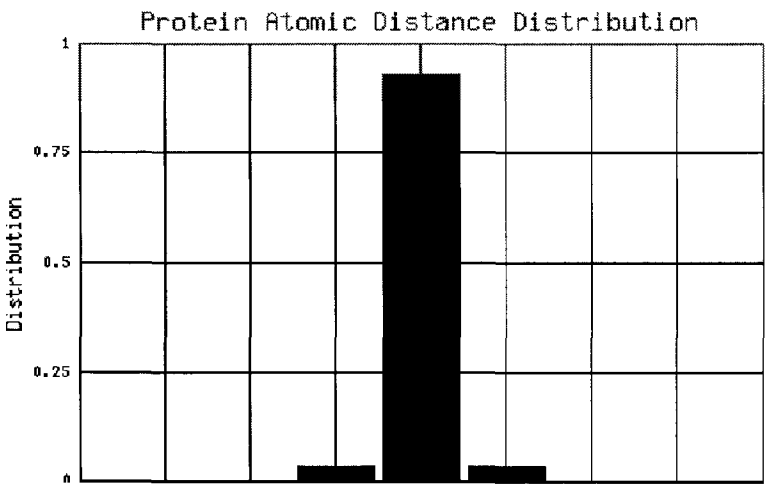
- This distance type is new and just has been studied!

This distance distribution has total 902 sampling distances from structural database of PIDD.

To obtain the distribution data, click

Download Distribution

Click to enlarge the size of the picture



Back to Top

APPENDIX D. SUBROUTINE OF MEAN FORCE POTENTIALS IN PROTEIN STRUCTURE REFINEMENT (IN FORTRAN 77)

In this subroutine, we compute the energy as well as the gradient.

If you need more information and implementation details, please write to the author Di Wu (diwu@iastate.edu)

```
--------------------------------
      SUBROUTINE APMF(E,TEMP5)
C
C Main target function routine
C
C Authors: Di Wu
C
C
C
C
      IMPLICIT NONE
C input/output
      INCLUDE 'cns.inc'
      INCLUDE 'comand.inc'
      INCLUDE 'coord.inc'
      INCLUDE 'deriv.inc'
      INCLUDE 'heap.inc'
      INCLUDE 'mtf.inc'
      INCLUDE 'cnst.inc'
      INCLUDE 'consta.inc'
      INCLUDE 'ener.inc'
      INCLUDE 'param.inc'
      INCLUDE 'timer.inc'
      INCLUDE 'funct.inc'
```

```
      DOUBLE PRECISION E,gama,gama1,TEMP5
C LOCAL


      INTEGER I,n7

      OPEN(49,FILE='testn.dat',STATUS='OLD')
      DO I=1,1
        READ(49,48) n7,gama,gama1
48    FORMAT(I7,F8.3,F10.2)
      ENDDO
      CLOSE(49)

      CALL APMF2(E,n7,gama,gama1,TEMP5)

      END


      SUBROUTINE  APMF2(E,n7,gama,gama1,TEMP5)

      IMPLICIT NONE

C input/output
      INCLUDE 'cns.inc'
      INCLUDE 'comand.inc'
      INCLUDE 'coord.inc'
      INCLUDE 'deriv.inc'
      INCLUDE 'heap.inc'
      INCLUDE 'mtf.inc'
      INCLUDE 'cnst.inc'
      INCLUDE 'consta.inc'
      INCLUDE 'ener.inc'
```

```
      INCLUDE 'param.inc'
      INCLUDE 'timer.inc'
      INCLUDE 'funct.inc'
      DOUBLE PRECISION E

C LOCAL
      INTEGER m,n,d,f,I,J,n7

      DOUBLE PRECISION g7(n7,3),esmall,gd1,gd2
      DOUBLE PRECISION gama,gama1,TTEMPD,TEMP5
      DOUBLE PRECISION a,b,c,d1,d2,d3,dt,prob
      INTEGER a7(n7,2)


      OPEN(50,FILE='test.dat',STATUS='OLD')

      m=0
      n=0

      DO I=1,n7
      READ(50,60) d,f,a,b,c
        m=m+1
        a7(m,1)=d
        a7(m,2)=f
        g7(m,1)=a
        g7(m,2)=b
        g7(m,3)=c
60    FORMAT(2I7,3F8.3)
      ENDDO
      CLOSE(50)

      TTEMPD=TEMP5+1.0
C       PRINT *, 'hello, be careful'
```

```
C     WRITE(*,*) TEMP8
C     WRITE(*,*) TTEMPD

      gama=gama*TTEMPD
C     PRINT *, gama

      DO 10 I=1,n7
      d1=(X(a7(I,1))-X(a7(I,2)))*(X(a7(I,1))-X(a7(I,2)))
      d2=(Y(a7(I,1))-Y(a7(I,2)))*(Y(a7(I,1))-Y(a7(I,2)))
      d3=(Z(a7(I,1))-Z(a7(I,2)))*(Z(a7(I,1))-Z(a7(I,2)))
      dt=sqrt(d1+d2+d3)

C     prob=g7(I,3)*exp(-(dt-g7(I,1))*(dt-g7(I,1))/(g7(I,2)*g7(I,2)))
      prob=-(dt-g7(I,1))*(dt-g7(I,1))/(g7(I,2)*g7(I,2))
C     PRINT *, prob

      esmall=-gama*(log(g7(I,3))+prob)

C     PRINT *, esmall
      E=E+esmall
      gd1=gama*2*(dt-g7(I,1))/(dt*g7(I,2)*g7(I,2))

C     PRINT *, gd1
      DX(a7(I,1))=DX(a7(I,1))+gd1*(X(a7(I,1))-X(a7(I,2)))
      DY(a7(I,1))=DY(a7(I,1))+gd1*(Y(a7(I,1))-Y(a7(I,2)))
      DZ(a7(I,1))=DZ(a7(I,1))+gd1*(Z(a7(I,1))-Z(a7(I,2)))

      DX(a7(I,2))=DX(a7(I,2))-gd1*(X(a7(I,1))-X(a7(I,2)))
      DY(a7(I,2))=DY(a7(I,2))-gd1*(Y(a7(I,1))-Y(a7(I,2)))
      DZ(a7(I,2))=DZ(a7(I,2))-gd1*(Z(a7(I,1))-Z(a7(I,2)))

10    CONTINUE
      END
```

APPENDIX E. REFINEMENT ON COMPARATIVE MODELS WITH MEAN FORCE POTENTIALS

Here we report some results on CASPR competition with using mean force potentials. As explained in the CASPR prediction center, it has become clear that refinement of comparative models of protein structures is a major challenge. Even though current protein structure prediction methods could provide some good initial structures or templates, the requirement of refining these structures is still important and necessary. We apply the developed refinement protocol with mean force potentials to these comparative models.

The methodology we introduced here actually combines solving distance geometry problems, molecular dynamic simulations and mean force potentials. First of all, we ran CHARMM in parallel on multi-processors and obtained a large set of energy minima based on the given target structure. Based on the energy and Ramachandran plot, we selected some typical structures. For each structure, we generated a set of distance bounds for it by allowing some distances to be flexible by 20%. And then using those generated distance constraints, we started CNS combined with mean force potentials to rebuild its ensembles. Again we use energy or Ramachandran plot to select the most possible structure. However, current criteria for selecting these structures are still problematic and also exist in our work. But a set of possible structures are still obtained and we report the one with the lowest rmsd value to the target structure.

The target structure we used here is downloaded from the website of CASPR, with PDB name 1WHZ which is hypothetical protein and has 70 residues. The target structure actually was predicted by the Baker group in CASP6. The rmsd value between the target structure and the true structure 1WHZ is 2.19 Å (see picture 31). It is easy to see that for beta-sheets the predicted structure does not have wiggles and some loop regions are also very different or poorly modeled, compared to the true structure. We used our refinement protocol described above to model the target structure. The structure has been further improved and the rmsd value was substantially reduced to 1.80A from 2.19A originally. Also the new refined model has also the same wiggles in those beta-sheet regions (see picture 32) as the true structure, while the original target structure does not have.

In this project, the mean force potentials and dynamic simulation indeed refine the structure and such conclusion could be obtained from the energy and rmsd values. However, in this initial application, the generation and selection might be still problematic. Especially, after dynamic simulation or structure determination, the energy and rmsd are not exactly highly correlated and it

hence brings a challenging question to refinement. It also implies the energy generated from current force fields is still not reliable and the requirement of using more accurate scoring function for selecting native structures is specifically important. However, in this project, we did have chance to investigate all these interesting topics.

Figure 34. The comparison of the predicted (a) and true structures (b) of 1WHZ

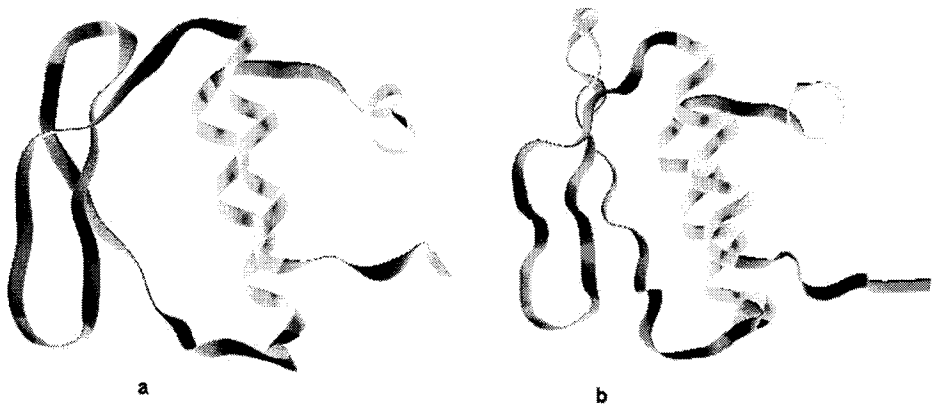
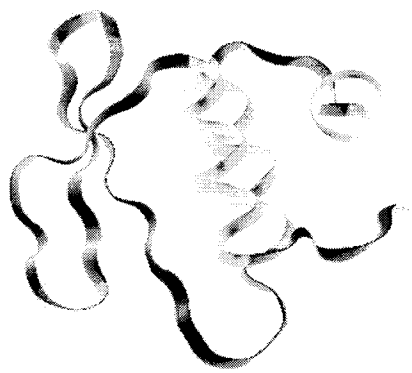


Figure 35. The refined target structure



The future direction of applying the mean force potentials to comparative model refinement will include the investigation on generating reliable structures, developing a more accurate scoring function for choosing the most native structures, and refining structures with additional mean force potentials. Such detailed report will be available later.

APPENDIX F. MATLAB CODE OF LOCAL-DME CALCULATIONS AND GASUSSION NETWORK MODEL

In this code, the main function includes the general Local-DME calculations and also there is a subroutine specifically for GNM calculations (GNMb=GNM(coord,n))
Please write to the author Di Wu(diwu@iastate.edu) for more information.

```
function dme_analysis(file1,file2)

av_m=fopen(file1,'r');
av_n=fopen(file2,'r');

n1=0;
%1 means NMR, 2 means Cry.
while(~feof(av_m))
   clear origindata;
   origindata=fgets(av_m);
   clear a1;
   a1=origindata(1,1:4);
   if strcmp(a1,'ATOM')
      clear a2;
      a2=origindata(1,14:15);
      a3=origindata(1,17);
      a4=origindata(1,22);
      if strcmp(a2,'CA')&(strcmp(a3,' ')|strcmp(a3,'A'))&(strcmp(a4,' ')|strcmp(a4,'A'))%|strcmp(a2,'N
')|strcmp(a2,'C ')|strcmp(a2,'O ')
         n1=n1+1;

coord1(n1,1)=str2num(origindata(1,31:38));coord1(n1,2)=str2num(origindata(1,39:46));coord1(n1,3)
=str2num(origindata(1,47:54));
      end
   end
end
```

```
fclose(av_m);

n2=0;
while(~feof(av_n))
   clear origindata;
   origindata=fgets(av_n);
   clear a1;
   a1=origindata(1,1:4);
   if strcmp(a1,'ATOM')
      clear a2;
      a2=origindata(1,14:15);
      a3=origindata(1,17);
      a4=origindata(1,22);
      if strcmp(a2,'CA')&(strcmp(a3,' ')|strcmp(a3,'A'))&(strcmp(a4,' ')|strcmp(a4,'A'))%|strcmp(a2,'N
')|strcmp(a2,'C ')|strcmp(a2,'O ')
         n2=n2+1;

coord2(n2,1)=str2num(origindata(1,31:38));coord2(n2,2)=str2num(origindata(1,39:46));coord2(n2,3)
=str2num(origindata(1,47:54));
         bfactor(n2,1)=str2num(origindata(1,61:66));
      end
   end
end
fclose(av_n);
n1
n2
m=n1/n2;

dmev=zeros(n2,1);
for i=1:m
   clear coorda;
   coorda(1:n2,1:3)=coord1((i-1)*n2+1:i*n2,1:3);
   for j=i+1:m
```

```
      clear coordb;
      coordb(1:n2,1:3)=coord1((j-1)*n2+1:j*n2,1:3);
      dmedata=dme(coorda,coordb,n2);
      dmev=dmev+dmedata;
   end
end

dmev=dmev*sum(bfactor)/sum(dmev);
GNMb=GNM(coord2,n2);
GNMb=GNMb*sum(bfactor)/sum(GNMb);
corrcoef(dmev,bfactor)
corrcoef(GNMb,bfactor)
corrcoef(GNMb,dmev)
bfactor
GNMb
dmev
function dmedata=dme(coord,coordt,n)
dmedata=zeros(n,1);

for i=1:n
   for j=1:n
      d1=norm(coord(i,1:3)-coord(j,1:3),2);d2=norm(coordt(i,1:3)-coordt(j,1:3),2);
      dmedata(i,1)=dmedata(i,1)+(d1-d2)*(d1-d2);
   end
end

dmedata=sqrt(dmedata);

function GNMb=GNM(coord,n)
contact=zeros(n,n);
coord
for i=1:n
   for j=1:n
```

```
      d1=norm(coord(i,1:3)-coord(j,1:3),2);
      if d1<7&i~=j

        contact(i,j)=-1;
      end
    end
end

for i=1:n
   for j=1:n
      if i~=j
         contact(i,i)=contact(i,i)-contact(i,j);
      end
   end
   contact(i,i);
end
contact;
[U,S,V]=svd(contact);

        B=zeros(n,n);
        for i=1:n
           Sv(i)=S(i,i);
        end
        [Ss,I]=sort(Sv);
        L=0;
        Ss

        for i=1:n
           if Ss(i)>1e-3
           %   L=L+1;
           B=B+U(:,I(i))*U(:,I(i))'/Ss(i);
           %end
           %if L==5
```

```
    %    break;
   end
end

for i=1:n
   Bpred(i,1)=B(i,i);
end
GNMb=Bpred/(sum(Bpred));
```

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, I would like to thank Dr. Zhijun Wu for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would like to thank Dr. Robert Jernigan for his great help and strong support during my graduate study. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Drena Dobbs, Dr. Kai-ming Ho and Dr. Vasant Honavar. I would additionally like to thank BCB program, department of mathematics and Baker center in Iowa State University for providing me the financial support and computational facilities. All my colleagues who have ever assisted me are gratefully acknowledged as well.